Statistical analysis for longitudinal MR imaging of dementia

Gerard Robert Ridgway

University College London

2009

Abstract

Serial Magnetic Resonance (MR) Imaging can reveal structural atrophy in the brains of subjects with neurodegenerative diseases such as Alzheimer's Disease (AD). Methods of computational neuroanatomy allow the detection of statistically significant patterns of brain change over time and/or over multiple subjects. The focus of this thesis is the development and application of statistical and supporting methodology for the analysis of three-dimensional brain imaging data. There is a particular emphasis on longitudinal data, though much of the statistical methodology is more general.

New methods of voxel-based morphometry (VBM) are developed for serial MR data, employing combinations of tissue segmentation and longitudinal non-rigid registration. The methods are evaluated using novel quantitative metrics based on simulated data. Contributions to general aspects of VBM are also made, and include a publication concerning guidelines for reporting VBM studies, and another examining an issue in the selection of which voxels to include in the statistical analysis mask for VBM of atrophic conditions.

Research is carried out into the statistical theory of permutation testing for application to multivariate general linear models, and is then used to build software for the analysis of multivariate deformation- and tensor-based morphometry data, efficiently correcting for the multiple comparison problem inherent in voxel-wise analysis of images. Monte Carlo simulation studies extend results available in the literature regarding the different strategies available for permutation testing in the presence of confounds.

Theoretical aspects of longitudinal deformation- and tensor-based morphometry are explored, such as the options for combining within- and between-subject deformation fields. Practical investigation of several different methods and variants is performed for a longitudinal AD study.

Declaration

I, Gerard Robert Ridgway, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgements

I must begin by thanking Derek Hill, not only for his supervision and direction, but also for his genuine enthusiasm and excitement for the field — it was primarily this that persuaded me to begin a PhD in the Centre for Medical Image Computing. Nick Fox has been an extremely patient and supportive supervisor and mentor; I could not have made it this far without his guidance and reassurance. I am especially grateful to Sébastien Ourselin, for accepting the unenviable task of taking over as my primary supervisor half way through the PhD and ensuring that the other 'half' took proportionally less time.

Brandon Whitcher and Tom Nichols have earnt my sincere gratitude, firstly through helping me immeasurably with my work, but also by inspiring me with the quality of their own research. Thanks to them and to Anil, Chris, Becky, Jonathan and others at GSK for making my placement there an enjoyable as well as productive experience. I am grateful to GSK and the EPSRC for the CASE studentship that supported me financially.

I've been very fortunate to have worked with some great people at CMIC, from whom I've learnt a lot. I'm particularly grateful to Oscar Camara, Bill Crum and Julia Schnabel. Danny Alexander did a brilliant job of examining my first and second year reports; his input has undoubtedly strengthened this thesis. To my official examiners, John Ashburner and Christian Beckmann, I'd like to say thanks in advance for reading through all of this, and I should clarify that it's not Danny's fault that it's so long! As I've neared the end of my PhD, I've particularly enjoyed trying to help the newer research students in CMIC, most notably Marc Modat, who has in fact helped me a great deal, both through expertly proofreading one of the most detailed sections of my thesis, and through the many interesting discussions we have had.

The Dementia Research Centre has been a truly lovely group with which to collaborate, and I am grateful to everyone there with whom I've had the pleasure of working. Rachael Scahill deserves special mention for providing guidance and support in the crucial early stages of my PhD, and for first teaching me about VBM. Chris Frost has been an indispensible source of assistance and guidance with respect to statistics, and has been admirably patient in helping an engineer to learn the finer points of an imaging-statistics approach in which Chris himself has no interest! It's been a real privilege to work with Jonathan Rohrer, Nicola Hobbs, Rohani Omar, Jason Warren, and many others. Susie Henley deserves special mention for being both a great friend and also crucial source of thesis-writing tips and support.

John Ashburner has tought me a great deal about the theory of computational anatomy and provided substantial practical help with VBM and SPM, for which I'm very grateful. I'd also like to thank other friends and colleagues at the FIL and elsewhere; in particular, I would like to say a special thank you to Klaas Enno Stephan, whose friendship and faith in me helped to restore my self-confidence at a time when I was particularly struggling with my work.

My parents have been wonderfully supportive throughout my PhD, as with everything, for which I am eternally grateful. My sister has done an excellent job of reassuring me that I would get there in the end; I hope I can be as supportive in return as she nears the end of her own thesis.

More than anyone else, my girlfriend has suffered through the low points of the PhD with me. Without her loving support and encouragement I would almost certainly have given up, and so — though she deserves a better (and a much more timely) one — this thesis is dedicated to Lucy Williams.

Contents

List of Figures

List of Tables

Chapter 1

Background

Scientifically interesting and clinically important patterns of longitudinal brain change in dementia may be detected through the analysis of serial MR imaging. Differences in these patterns between groups of subjects may be relevant to diagnosis, tracking of disease progression, and monitoring of potential disease-modifying treatments.

This thesis focuses on the application of techniques from the fields of image analysis and statistics to the problem of identifying such patterns of change and their inter-group differences. This chapter presents some of the clinical background, followed with brief introductions to the key image analysis techniques, and basic statistical methods.

1.1 Clinical Application

The term dementia can refer to a number of different neurological disorders which result in impaired functioning of the brain. Reasoning, memory, emotional behaviour, speech, and movement, are among the faculties that can be degraded. Dementias include Alzheimer's Disease, Vascular Dementia (caused by problems with the brain's blood supply), Lewy Body Dementia (which shares certain traits of Alzheimer's and Parkinson's Disease), Fronto-Temporal Dementia (which affects emotional judgement and social behaviour), Huntington's Disease (an hereditary disorder affecting personality and movement), and Creutzfeldt-Jacob Disease (a rare and rapidly fatal encephalopathy) [1, 2].

Alzheimer's Disease (AD) is the most common form of dementia. It is a progressive disorder that leads to problems with memory, learning, judgement, communication, and the basic abilities needed for independent living. At present, researchers know of no single cause, nor of a cure. The average life-expectancy after first symptoms is eight years, though this varies widely [3].

The greatest risk factor for AD is age — about a tenth of people older than 65 are affected, rising to almost half of those over 85 [4].[1] In many of the world's countries the average age of the population as a whole is increasing, and the health and economic burden of AD will grow as a consequence. Ferri et al. [5] estimate that over 24 million people suffer from dementia, world-wide, and they predict this will rise to over 80 million

[1] Prevalence may vary geographically; these figures were estimated from the United States population.

by 2040. The financial impact of AD is severe, both in terms of patient care and lost workforce productivity, and is motivating significant research investment.

Diagnosis of dementia chiefly involves clinical examination and consideration of medical history. Diagnostic tests include the Mini-Mental State Examination [6] and the NINCDS-ADRDA criteria [7]. Structural Magnetic Resonance Imaging can play an important part in early diagnosis, tracking of disease progression, and differentiation of dementia from other diseases [8, 9, 10, 11, 12]. AD can only be proven by post-mortem histopathological examination.

People found to have memory problems incommensurate with normal ageing, but without sufficient cognitive problems to fulfil criteria for dementia, are said to have Mild Cognitive Impairment (MCI) [3]. Some such people later go on to suffer from full AD, while others appear not to; the conversion rate is approximately 10–15% per year, which is around ten times higher than the rate in the healthy elderly [13]. There has been some debate as to whether MCI is a distinct state, or simply an early stage of AD [13, 14]. Little is known about the neurological differences between MCI and AD (or the conversion process, if such an interpretation is correct), and this is a key area in which imaging research may offer vital clues [15].

Family history of AD is a risk factor — individuals with an affected parent or sibling are two to three times more likely to develop the disease [3], though genes are rarely the direct cause of AD. Distinction is made between Sporadic and Familial forms of AD [16]. Sporadic AD refers to the common, unpredictable, form of the disease, which is not caused by any particular gene. There are, however, genetic mutations which can increase the risk of developing the disease, and lower the typical age of onset. Familial AD is directly caused by an autosomal-dominantly inherited mutation in one of three (currently known) genes,[2] which almost invariably results in disease development and usually causes a much earlier onset (typically before the age of 60) [3]. Due to the predictability of Familial AD it has been possible to study at-risk individuals from a presymptomatic stage [8, 11], allowing important observations to be made about the very early progression of the disease.

While the underlying cause of AD is still unknown, the histopathological hallmarks were first noted 100 years ago by Alois Alzheimer, and have since been extensively characterised [17]. Two distinct features can be identified: *Neurofibrillary tangles* are twisted strands of abnormal Tau protein that form within the neuronal cells; *Neuritic plaques* are extra-cellular clumps built up from beta-amyloid (Aβ) — a fragment of a protein (APP). Because the amyloid plaques are much larger and form outside cells, there is potential for imaging them using high field strength MRI [18], though such techniques have thus far only been demonstrated *ex vivo* on autopsy specimens [19] and *in vivo* in mice [20]. Research is also progressing on MR imaging of amyloid using labelling compounds [21], PET imaging with the marker molecule PIB [22, 23], and Near-Infrared imaging [24]. Amyloid (and Tau) can also be detected in CSF via lumbar puncture [23, 25].

Some researches believe that the accumulation of Aβ is the primary *cause* of AD, rather than just being symptomatic [26]. There is controversy regarding this 'Amyloid

[2] All of which relate to the production of beta-amyloid — see later.

Hypothesis' and authors dispute the exact mechanism by which it might occur [27]. Nevertheless, there has been great interest in the possibility of treating AD by targeting Aβ, either directly, or by inhibiting the processes that led to its creation [28] or its deposition as plaques [29]. Schenk et al. [30] showed that immunization with Aβ attenuated pathology in a mouse model of the disease, and a trial was later conducted on humans [31]. The trial was prematurely terminated due to meningoencephalitis in some subjects, but results indicated some slight cognitive improvement in antibody responders, and lumbar puncture samples showed reduced Tau present in CSF [31]. Cases that have come to post-mortem showed immunoreactive macrophages and the absence of plaques in certain regions [32]. Imaging results from the trial showed, surprisingly, that drug responders had *increased* rates of whole-brain and hippocampal atrophy and ventricular expansion [33], going against the prior expectation that the drug would slow disease progression and hence reduce customary measures of atrophy. The apparent greater atrophy might however indicate clearance of Aβ, so there remains hope that the drug is beneficially modifying the disease. Investigation of local regional patterns of atrophy in these subjects, and in particular the differences between atrophic patterns in placebo and drug responder groups, may help lead to important deductions about the exact effect of the drug.

1.1.1 Summary of Clinical Questions

The following is a brief list of some interesting clinically-focussed research questions in the field of dementia, which may be usefully addressed with serial imaging-based studies, particularly in terms of group differences in longitudinal patterns of atrophy:

- How does AD differ from normal ageing?
- What structural neural changes distinguish AD from MCI?
- What distinguishes MCI subjects who progress rapidly to AD from those who don't?
- Are there structural/atrophic differences between Familial and Sporadic AD?
- Does the rate or pattern of atrophy differ when drugs are administered?
- How does atrophy in less common dementias compare to that in AD?

1.2 Overview of Methods

An overwhelming range of image analysis methods have been applied to neuroimaging, many of them applicable to imaging of dementia, many suitable for longitudinal statistical analysis, and some of them specifically tailored to serial data. One potential broad taxonomy of methods would be to divide them into: those that directly measure some quantity from an image, or a change between two images; those that analyse intensities, segmentation probabilities, or registration transformations, on a voxel-wise basis; and those that study surfaces extracted from the data. Methods will be discussed briefly in this general order below.

The algorithmically simplest methods involve large amounts of manual interaction. For example, one of the best known examples of this type of method is to manually (or semi-automatically) segment anatomical regions of interest [34], after which their volumes or other properties may be quantified and statistically compared. Another example is shape analysis based on manually selected point landmarks and their correspondences across sets of images [35].

Two more complicated approaches, designed specifically for pair-wise longitudinal imaging of brain change, involve rigid or affine registration of the image pair, followed by estimation of the movement of the brain/CSF boundary. The Boundary Shift Integral (BSI) [36] estimates the volume-change 'swept out' by the moving boundary between high and low intensity regions by integrating the resultant intensity differences caused by the movement. SIENA estimates the movement of the boundary by analysing the correlation of the intensity profiles normal to the boundary in both images [37].

These techniques give useful and well-validated measures of volume change due to atrophy in dementia. However, they are mainly useful for whole-brain analysis, with some possibility for regional application if combined with manual segmentation of a baseline region. Many of the clinical questions discussed above can be better answered if a finer scale assessment of local atrophy in many different regions of the brain can be determined. Secondly, there is a need to be able to compare regional effects across subjects, which requires some form of datum or standard space (something discussed further in section 1.5.1).

These desires motivated the development of Voxel-Based Morphometry (VBM) [38, 39, 40], which aims to give a whole-brain voxel-wise analysis of morphological differences between groups. VBM basically involves segmentation, non-rigid registration or 'normalisation' into the standard space of a template image, smoothing (for both theoretical and practical reasons), and voxel-wise statistical testing to determine which areas differ significantly over time or between groups. The technique of VBM has gained very wide use, especially in the fields of ageing and dementia [15, 41, 42, 43, 44], though its basic underlying philosophy has been challenged by some authors [45, 46]. (It should be noted here that SIENA has also been extended for voxel-wise comparison in a related but distinct way [47].)

The original VBM approach looked at 'mesoscopic' differences remaining in the segmentations after the spatial normalisation. It was later altered to incorporate a 'modulation' step in which the volume change arising from the non-rigid registration was applied to the normalised segmentations by multiplying voxels by the determinant of their corresponding transformation Jacobian. For longitudinal analysis within a single subject, the Jacobians of a non-rigid registration displacement field can be directly studied in an approach known as Voxel-Compression Mapping (VCM) [11]. However, as mentioned earlier, comparisons between subjects require some further form of inter-subject registration to ensure approximate voxel-wise correspondence. There have been several approaches that combine non-rigid intra-subject registration of serial images with non-rigid inter-subject normalisation to standard space (usually known as Deformation- or Tensor-Based Mor-

phometry) [46, 48, 49, 50]. It should also be noted that the displacement fields themselves, or some other tensor measure can also be studied in place of the Jacobian determinants [39]. Much more on this will be included in later chapters.

Due to the clinical importance of the cerebral cortex in dementias including AD, methods have been developed which directly study an estimated cortical surface extracted from the data [51, 52, 53, 54, 55, 56]. As with voxel-based methods, the problem of correspondence for group-wise comparison is of key importance, and has been addressed in the references just given. Taking full advantage of longitudinal imaging data in surface based approaches has arguably received less attention than in the voxel-based literature, though successful longitudinal studies have been performed [57, 58]. Closely related approaches to the problem of surface extraction and modelling, for example using deformable shape models, have also been applied to structures other than the cortex, such as the hippocampus [59].

It is apparent that voxel-based and surface-based techniques are dependent on the more fundamental image analysis techniques of registration (especially non-rigid, for inter-subject comparison) and segmentation (either for tissue maps or as part of the surface extraction process), and these will be reviewed further in section 1.5. The importance of appropriate statistical analysis is also clear, particularly with regard to data with both longitudinal and cross-sectional aspects, and this will be discussed in some detail in section 1.6.

1.3 Image Acquisition

Magnetic Resonance Imaging (MRI) [60] uses the quantum-mechanical magnetic properties of nuclei (for example hydrogen, found in water in the cells of the body) to produce spectroscopic measurements or tomographic images. Through the application of magnetic fields and the transmission and reception of radio-frequency electromagnetic pulses, it is possible to sample the spatial Fourier domain (or 'k-space') signal of the object being imaged; the inverse Fourier transform can then be used to recover the spatial domain image.

The majority of the work presented in this thesis uses three-dimensional T1-weighted structural scans, acquired at a main field strength of 1.5 T, using a spoiled gradient echo sequence. Details of the acquisition process and the physics underlying the many alternative sequences [61, 62] will not be given here, though a brief look at some of the potential — from an image analysis perspective — of different MRI modalities will be taken in section 1.3.2.

First, some important aspects of the variability and reliability of MR imaging are reviewed, particularly with regard to possible changes over time. It should also be noted that images acquired on different scanners are likely to differ significantly in terms of intensities and geometrical accuracy, even if care is taken to use the same pulse-sequence and protocol. This issue is particularly important for large cohort studies (where multiple geographic locations may be required to recruit sufficient numbers), for clinical trials

(where multiple centres are desirable for statistical as well as practical reasons), and for long-term studies (in which scanners may be serviced, upgraded, or even replaced over the time-scale of the investigation).

1.3.1 Geometrical Distortion and Serial Imaging Stability

The k-space MR signal is sampled by means of spatial encoding, which essentially involves establishing spatial patterns of frequency and phase of the rate of precession of the macroscopic bulk spin of the voxels. The signal from a voxel depends on the resonant frequency of the atoms making up its tissue, and on the magnetic field strength at that point. The different resonant frequencies ('chemical shift') of water and fat (205 Hz difference, at 1.5 T) lead to a relative shift of their spatial domain signals [63]. The effect can be reduced with fat-suppressing acquisition sequences.

The magnetic field at a point depends on the main B_0 field, the applied gradient fields, and the magnetic susceptibility of the surrounding material. The main field should ideally be constant throughout the material; a process known as 'shimming' attempts to ensure this, both by the installation of fixed shims and by an electromagnetic 'auto-shimming' procedure which attempts to correct for material-induced magnetic susceptibility inhomogeneities at the time of each scan. Imperfections in this process (either from poor servicing, RF-heating of the shims in the course of scanning, or challenging levels of susceptibility variations) are one source of geometrical distortion. Eddy currents may also be induced in the main field windings by the gradient fields, altering the main field slightly (though the effect should be negligible in shielded-gradient systems) [63].

Non-uniformity in the applied magnetic gradients, caused by nonlinearity of the gradient coils, are one of the most significant sources of geometrical distortion. The coils are manufactured to ensure accuracy near the iso-centre of the main field; further away from this, the field nonlinearities become worse (Fig. 1.1 illustrates this effect). In serial imaging, inconsistent positioning of the patient between scans can lead to different nonlinear geometrical distortion being present in the images at different time-points — something that will confound the detection of biologically meaningful nonlinear changes of anatomical geometry (e.g. due to atrophy).

The problem of geometrical distortion in MRI has been recognised for a long time [64, 65], and many methods have been developed for its characterisation and correction. Sumanaweera et al. state that the gradient nonlinearities and magnetic susceptibility inhomogeneity are the two largest sources of distortion [63]. Gradient nonlinearities can be characterised quite accurately by the scanner manufactures, using for example a spherical harmonic expansion [66], and a method to correct the images based on this knowledge has been patented as 'GradWarp' [67].

At the present time, many clinical and research scanners are not properly corrected for the gradient nonlinearities, and scanner manufacturers do not always make the gradient field modelling information available [68]. It may also be the case that even after correction, some gradient-based geometrical distortion remains, perhaps due to the slight deviations in the scanner's coil performance from the specifications. This has motivated the development

of several phantom-based methods for characterising and correcting distortion, [69, 70, 71, 72, 73], including a comprehensive body of work by Wang et al. [74, 75, 76]. The image in Fig. 1.1 shows a typical phantom for this purpose.

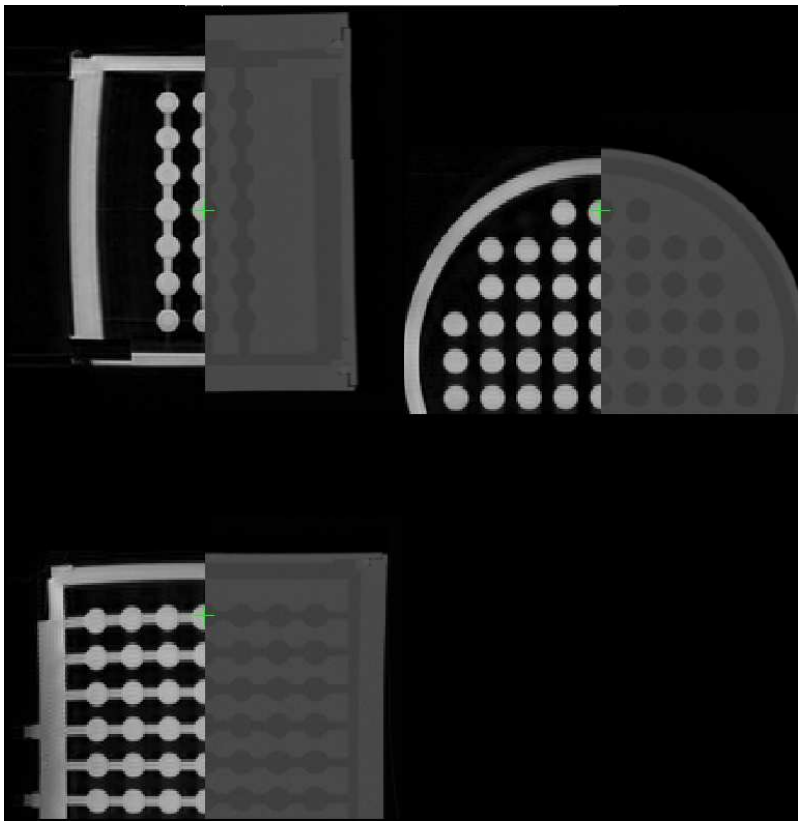


Figure 1.1: A phantom image, showing MR (left) with some distortion due to gradient nonlinearities apparent toward the edges, and a rigidly registered CT image (right).

Hill et al. investigated distortion in MR and CT using bone-implanted markers, and found that the Fiducial Registration Error for CT was below 0.5 mm [77]. This suggests the possibility of using MR-CT registration methods to correct MR distortion, bypassing the need for a physical model of the phantom. Other authors, however, have found similar magnitudes of distortion in more modern MR and CT images [78].

Because magnetic susceptibility variations depend on the material being imaged, they cannot be corrected by phantom methods. Two different approaches have been popular: Chang & Fitzpatrick pioneered a technique using read-out gradient reversal [79, 80, 81, 82]; and Sumanaweera et al. developed a phase-mapping approach [63].

The validity and statistical power of serial imaging is not only affected by geometrical distortion, but also by other stability issues, such as uninteresting biological variability between time points. For example, dehydration or other short-term physiological conditions may have a sufficiently large effect on the brain to confound the measurement of small atrophic changes (a particular problem for short-interval studies), as reported by Littmann et al. [73]. The Alzheimer's Disease Neuroimaging Initiative[3] are also exploring such issues [83].

[3] www.alzheimers.org/ADNI.

1.3.2 Potential of other Imaging Modalities

The T1-weighted volumes used throughout this thesis are the most popular general purpose 3D MR acquisition method; they can provide good contrast between grey and white matter, and a good balance between resolution, signal-to-noise, and scanning time. Most scanners in clinical use today have 1.5 tesla fields, though 3 T scanners are becoming more common. The doubling in field strength can be used to double the SNR, or it can be shared between improvements to SNR, resolution, and scanning speed.

T2-weighted imaging allows vascular lesions in the white matter to be visualised and quantified [84], and research into their relationship with ageing and dementia continues to be carried out [85, 86]. Wen et al. have investigated the correlation between reduced grey matter volume (using T1 images) and increased volume of white matter lesions (using T2 FLAIR) [87].

Multi-modal T1 and T2 (and possibly further additions such as proton density) images of the same patient can be registered together using appropriate similarity measures, such as mutual information [88, 89, 90]. Because they have different tissue-contrast properties, the fused T1+T2 data contains more information to aid brain tissue segmentation, and algorithms exist for such multi-spectral segmentation [91, 92, 93].

Magnetisation Transfer Ratio (MTR) imaging measures differences between fixed and free protons (e.g. within cell walls vs. within fluid), and can reflect underlying microscopic pathological changes [94]. It has been applied in studies of ageing and dementia [95, 96].

Perfusion imaging measures the amounts of blood delivered to different parts of the brain. Hayasaka et al. have applied the technique to AD, and have also investigated the multi-modal fusion of information from conventional structural imaging and perfusion [97].

Diffusion imaging measures the mobility of water in tissue, and by considering the relative mobility in different directions Diffusion Tensor Imaging can be used to make inferences about the white matter tract connectivity. Rose et al. used DTI and found significantly reduced white matter tract integrity in AD compared to controls [98]. Both perfusion and diffusion-weighted MRI have been investigated in [99]. DTI and MTR were studied in [100] with application to AD; and both were analysed alongside conventional imaging in [101], with a focus on healthy ageing.

MR Spectroscopy can be used to measure metabolites including neuronal markers such as N-acetyl aspartate (NAA), which may indicate cell viability. This approach has been investigated in the field of neurological disorders including AD [12, 102, 103, 104].

Future developments in higher field strength imaging will allow higher resolution, and higher SNR imaging, such developments may be combined with the use of surface coils to focus on the cerebral cortex [105], potentially permitting much more accurate measurement of changes that occur in dementias, such as the cortical thinning that takes place in AD. Such improvements may also lead to the feasibility of applying sophisticated measures of shape change such as curvature of the cerebral cortex [106] to AD.

1.4 Clinical Validation

If methodological image analysis techniques are to be of genuine use in the domain of medicine then they must be carefully validated. The underlying assumptions, limits of applicability, likely error magnitudes, and possible failure mechanisms should all be investigated before techniques are relied upon for critical decisions such as those affecting patient care or the declared efficacy of a potential drug.

Validation can involve statistical or machine learning analyses, such as consideration of group-separation or the generalisation performance from optimising the algorithms on one data-set and then applying them blindly to another. It can also involve, for example, comparison between measures of disease progression with the known natural history of the disease, or comparison of regional measurements with known regional distributions of pathology.

As mentioned earlier, the exact nature of a clinically observed dementia can only be fully understood through post-mortem examination of histopathology. This motivates the validation of MRI measures and algorithms by comparing them with histological data, including images of stained tissue samples [107, 108].

Research has been carried out on the construction of volumes from multiple digitised photographic images of histology [109, 110] and the registration of such images with post-mortem MRI and with *in vivo* MRI

Fig. 1.2 shows an example of a simple approximate registration of MR volumes of fixed brain slices with the fixed (but unsliced) hemisphere from which they came. More advanced methods along similar lines should enable registration of digital photographs of stained histological blocks into the space of the fixed hemisphere MR image. If the fixed hemisphere can be successfully warped to the *in situ* post-mortem brain, and then back to available *in vivo* serial scans, then the photographic histology can be directly compared with any algorithmic measurements derived from the longitudinal MRI data.

1.5 Image analysis overview

1.5.1 Image Registration

The process of image registration basically involves the estimation of geometrical transformations that align two or more images. A registration algorithm requires a model for the class of transformation allowed, some measure of the quality of the alignment, and some means of optimising the model parameters to achieve a good solution. A suitable scheme for interpolation is also required. General overviews of medical image registration can be found in [111, 112, 113].

Transformation Models

Models for the geometry of the transformations can be categorised in several different ways, one broad classification divides them into affine transformations (sometimes inaccurately

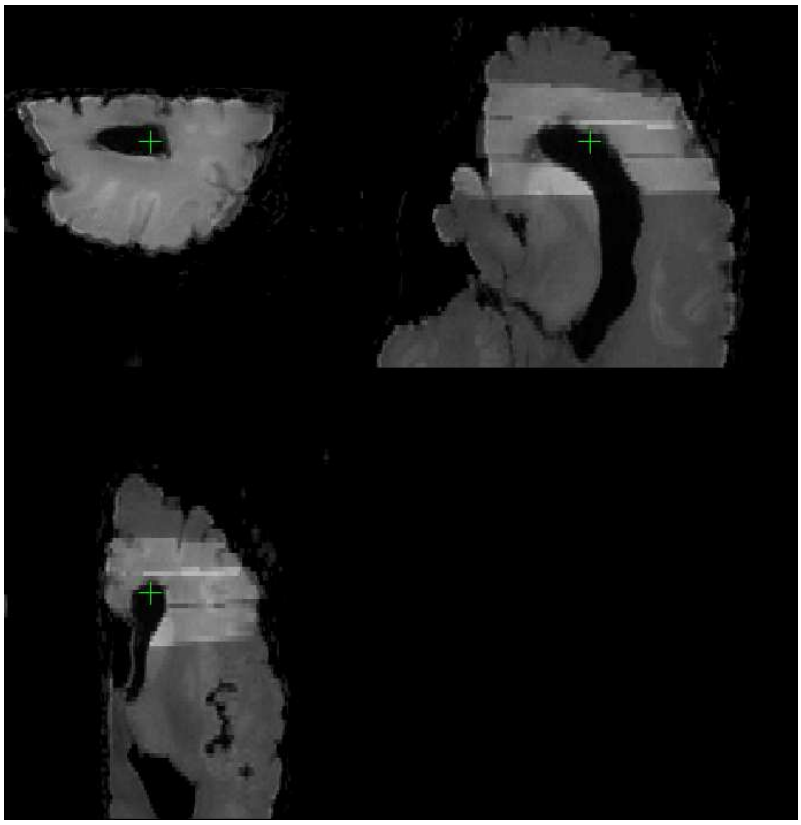


Figure 1.2: An illustration of the registration of three chemically fixed brain slices back to their parent hemisphere (shown overlaid, the slices appear bright). The programs rview and pareg from the I(R)TK software library of Daniel Rueckert and Julia Schnabel were used in this work.

referred to as 'rigid') and more general transformations which allow flexible distortion of the images (usually called 'non-rigid') [112, 114].

Affine transformations in three-dimensional space have 12 degrees of freedom (DF), which can be parametrised in several different ways, including: a general 3-by-3 linear mapping (9 DF) of coordinates with an additional translation (3 DF); three skews of the three axis pairs (e.g. altering the angle between the x- and y-axis from 90°), three re-scalings along each axis (either before or after they are skewed), a general rotation (3 DF), and a translation; or, using the polar decomposition analogous with principal strain in the mechanics of solids (see section 4.2.5), a rotation, three scalings along the rotated axes, a further rotation, and a translation. A general three-dimensional rotation has three DF, which can be parametrised in various ways, including three separate rotations about the Euler-angles, a rotation about an axis defined by a unit vector, or equivalently, with a unit quaternion.

Simple transformations involving combinations of one or more of: rotations, translations, scalings, and skews, are all special cases of affine transform. Popular cases include: rigid transformations with just rotation and translation (6 DF), which preserve lengths and angles; rigid with isotropic scaling (7 DF), which preserves angles and ratios of lengths;

and rigid with axis-aligned anisotropic scaling (9 DF).[4] Affine transformations preserve straightness and parallelism of lines. The combination of successive affine transformations is also affine.

Non-rigid models include polynomial mappings of the coordinates [115], and more complicated (high DF) modelling of 3D displacement fields. A distinction can be drawn between models which parsimoniously parametrise the displacement field, and those that parametrise the full voxel-wise 3D displacement but regularise the optimisation using appropriate physics-based constraints. Examples of the former include popular techniques using tensor products of cubic B-splines [116] or discrete cosine transform basis functions [117]. The latter have a long history, and include many types of physical deformation model, such as elastic [118], fluid [119, 120, 121, 122], anisotropic diffusion [123], and others. Details can be found in a recent book [124]. There is some overlap between these classes, since lower DF models such as free-form deformations [116] may also use physics-based regularisation, such as including elastic bending energy in the cost function.

Diffeomorphic mappings

It is common to assume that registration algorithms should recover physically reasonable deformations, having positive Jacobian determinants everywhere (see section 4.2). Such transformations are invertible, and the inverse mapping also has positive Jacobian determinants (the Jacobian matrix of the inverse mapping at the point mapped to, is the inverse of the Jacobian matrix of the forward mapping at the original point). If the domain and range are smooth (differentiable) manifolds and the transformation and its inverse are smooth (which implies that the Jacobian is bijective everywhere) then the mapping is a diffeomorphism. Diffeomorphisms form a mathematical group [125]; the practical relevance of this fact is that (a) the composition of two diffeomorphisms is another diffeomorphism, (b) the composition operation is associative, (c) there exists an inverse for each diffeomorphism, such that the composition of the pair returns the identity diffeomorphism. Furthermore, the diffeomorphism group can be given the structure of an infinite-dimensional Lie group, with associated exponential map and Lie algebra [125]. The exponential map means that the elements of the diffeomorphism group can be represented in the tangent plane, for example, diffeomorphic transformation fields can be produced by integrating sufficiently smooth velocity fields, which may be constant [126, 127] or variable [128]. Variable velocity fields can be derived from constant momentum maps [129]. Diffeomorphisms can also be found by solving the Euler equations on the group [130].

Elements in the diffeomorphism group (or their discretised approximation) cannot be treated as elements in a Euclidean space; addition or subtraction of displacement fields associated with diffeomorphic transformations will not necessarily yield diffeomorphic results. Similarly, concepts of distances between transformations and averages of transformations must account for the nature of the group. In essence, distances must be measured

[4] This class of transformation does not form a group, in that two successive 9 DF transformations can in general give a 12 DF affine transformation, and the inverse of a 9 DF transformation (though necessarily still 9 DF) can not generally be represented by the same nine parameters.

along curved geodesics along the surface of a Riemannian manifold. The concept of the mean must be generalised from the simple arithmetic case, through the principle that the mean minimises the average squared distance from itself to all other observations; using the Riemannian distance metric instead of the simple Euclidean one leads to the Fréchet mean [131]. One of the key advantages of the diffeomorphic approach is that the distance between transformations provided by the metric is also a natural distance between subjects [132, 133], upon which statistical or image-classification techniques can be based. Similarly, the distance between subjects can be important in terms of choosing a template [134].

Throughout this thesis, we use a simpler registration algorithm [135] which recovers continuous one-to-one mappings that may be composed and inverted, though they are not necessarily diffeomorphic, and nor can they represent any arbitrary diffeomorphism. The longitudinal focus here also means that the geodesic distance between subjects, and the related concept of a Fréchet mean [131], are not immediately applicable. This would be an important direction for further investigation.

Similarity Measures

Objective functions for alignment can be divided into two broad classes, those which use distance between point landmarks [35, 136, 137, 138], and those which consider some voxel-wise similarity measure. The latter include basic measures such as mean squared error and cross-correlation, and more sophisticated approaches (suitable for multi-modal registration, where a simple intensity relationship cannot be assumed). Popular and successful multi-modal similarity measures have been derived from information theory; they include joint entropy, mutual information [88, 139], and normalised mutual information [90]. These are reviewed in [140], contrasted together and compared to conventional measures in [141], and discussed in a non-rigid registration context in [142]. Another multi-modal similarity measure is the correlation ratio [89]. Roche et al. [143] have used a unifying mathematical approach to derive measures including cross-correlation, mutual information, and the correlation ratio, using maximum likelihood estimation.

Some more unusual similarity measures include examples involving the frequency domain [144], the images' local phase [145], or local frequency [146]. The use of image intensity gradients (in addition to other similarity measures) has also been investigated [147]. Authors have also derived other (non mutual information based) information theoretic measures [148, 149].

For non-rigid registration, landmark-based measures can give very reliable registrations of the chosen points (with accuracy determined by the quality of the point selection and any regularisation included) but may not achieve good global correspondence without huge numbers of landmarks being selected — a time consuming and error-prone process. Regarding voxel-based measures, the problems of local optima in the objective function, inadequate optimisation methods, and undesirable global optima that result in high voxel-wise similarity but anatomically poor correspondence or implausible deformations, can cause voxel-based registration to fail to meet clinically acceptable standards. This has

motivated the development of combined voxel-similarity and landmark-correspondence registration schemes. Landmarks may include manually selected points [150, 151, 152, 153], automatically identified regions or features [154, 155, 156, 157, 158]. Voxel-based registration can also be guided with manually or automatically located surfaces [56, 152, 159] or corresponding surface-based landmarks [160, 161, 162].

Some Extensions

A popular extension to rigid or non-rigid registration algorithms, designed to increase speed or robustness or both, is the implementation of a multi-level, multi-resolution, or hierarchical scale-space framework [122, 163, 164, 165].

High-DF registration algorithms can be very slow, and several avenues have been explored for speeding up the optimisation, including: advanced numerical approaches such as variational methods [166, 167]; parallel implementations [168]; computationally efficient strategies [169]; or modifications to the objective function [170].

A recent approach [171] using level-set methods [172], claims to unify some of the more conventional techniques.

A desirable extension is to include information available in data-sets containing more than two longitudinal images; some work on this has been done recently, but it is expected that more research in this area would be worthwhile. The CLASSIC algorithm of Xue et al. [173] aims for greater stability of longitudinal segmentation, by combining elastic warping and image adaptive clustering, using spatiotemporal smoothness constraints in joint estimation of the warps and fuzzy tissue segmentations.

Validation, Correspondence, and Atlases

As discussed in section 1.4, it is especially important that computational methods intended for clinical or medical research use are thoroughly tested and properly validated against specified standards. This is a challenging task for image registration, particularly non-rigid methods, as it is difficult to quantify the performance of an algorithm, or even the anatomical correspondences desired to result from it.

Several rigid intra-subject registration techniques were thoroughly compared by West et al. [174] using bone-implanted markers (hidden from the registration algorithms). Such an approach can easily give meaningful quantitative measures of registration quality, derived from the mismatch of known and registered marker positions. However, for non-rigid registration, bone-implanted markers cease to be useful, since the registration algorithms are able to locally redistribute tissue within the brain. Schnabel et al. [175] have used finite element simulations to provide a ground-truth deformation that can be compared with the results of non-rigid registration. In the case of neuroimaging of dementia, the changes that occur due to atrophy can be quite complex, so the generation of realistic simulated data is a difficult problem, and a current research topic.

For inter-subject registration (either rigid or non-rigid) implanted markers can't be used since their between subject correspondence would not be known. Crum et al. [176] have discussed the fundamental difficulty of deciding on correspondence between

the brains of different subjects (which can have large structural and functional differences in anatomy). Further more, common MR-visible borders and landmarks such as sulci do not necessarily relate precisely to more fundamental brain structure in terms of cytoarchitectonics [177].

An attempt has been made to 'evaluate,' though not fully *validate* — their distinction — non-rigid inter-subject registration in [178]. Their approach involved both global measures (such as overlap of tissue segmentations) and local measures (such as matching of selected sulci) and indicated that more sophisticated non-rigid registration techniques which attain superior global matching may actually be no better than rigid registration in terms of local anatomical correspondence. A similar study comparing two non-rigid methods can be found in [179].

Another method that may be used for evaluation of inter-subject non-rigid registration is to consider 'round-trip' errors in a sequence of registrations, e.g. register A-to-B then, separately, B-to-C, and C-A, and check the effect of following a full A-to-A loop. Indeed, many algorithms are not guaranteed to find consistent pair-wise transformations for A-to-B and B-to-A — in contrast with e.g. [180]. Crum et al. [181, 182] have established metrics that can be used to assess non-rigid and group-wise registration across an ensemble of images.

This issue provides a close link to another topic in the broad area of inter-subject comparison and the voxel-wise correspondence problem — the aim of template or atlas building. Population templates should ideally not be biased toward any particular image or sub-group of images. This has motivated several approaches specifically designed for group-wise registration or atlas construction. Kochunov et al. [183, 184] determine the average deformation field from a number of registrations to a target (possibly selecting the Best Initial Target on the basis of prior group-wise registration) and use this to transform the target to the Minimal Deformation Template. Other authors attempt to directly create an unbiased population atlas [126, 185, 186, 187, 188].

1.5.2 Tissue Segmentation

Segmentation of the human cerebral cortex — so important for some of the most popular techniques such as VBM and surface-based analyses — is particularly challenging, due to the complexity of the shape of the cortex. Image segmentation is a very widely researched topic; some classical methods were reviewed in [189]. Some popular general methods include: seeded region growing [190]; the watershed method [191]; active contours, or active shape models [59, 192, 193, 194, 195]; segmentation propagation [196, 197]; or combinations of these approaches.

Intensity-histogram based clustering methods such as k-means and fuzzy c-means [198] are well-suited to the segmentation of unpredictably-shaped regions. Other clustering methods include mean-shift [199], and recent spatio-temporal models [200]. In addition to clustering, full statistical modelling of the intensity histogram is also possible (Choi et al. [91] seems to be one of the earliest references).

Segmentation propagation is the process of using non-rigid registration of an image

which has been segmented (perhaps manually) with the image to be segmented and transferring the tissue labels across with the determined transformation field [196]. The propagated labels can either be treated directly as the segmentation of the new image, or they may be used as spatial priors in a combined registration and histogram approach, as implemented in several popular techniques [92, 201, 202].

Dependencies between neighbouring voxels (for example those that belong to the same tissue type) can be modelled to some extent using the theory of Markov Random Fields (leading to Gibbs Distribution priors), or Hidden MRFs [93, 202].

Too many techniques have been developed to discuss in detail here; some of the most cited relevant references include [203, 204, 205, 206, 207]. Some more unusual approaches include: the use of Support Vector Machine classifiers [208]; variational methods [209]; level-set based methods [172]; and the explicit modelling of longitudinal data [173].

Intensity-based segmentation can be degraded by the presence of intensity variations due to MRI RF field inhomogeneities. The task of correcting these bias fields is therefore a closely related and important topic. A review can be found within section 5.2.

Several tissue classification procedures require, or benefit from, the prior removal of tissue such as scalp, eyes, and neck; a process generally referred to as 'brain extraction' [210] (in this context, CSF is often considered brain tissue). Several such methods are reviewed in [211]; one of the most successful is found to be that described in [206].

As stated several times elsewhere in this chapter, it is crucial to consider the problem of validation. Since expert manual segmentations show both inter- and intra-observer variability, genuine 'ground-truth' for segmentation can rarely be known; acceptable 'gold-standards' for validation are available on data that has been simulated in some way, for example by using non-rigid registration of a well-labelled template to different subjects (keeping the transformed template and labels and not the original subject images). Various overlap measures can then be used to quantify the performance of different algorithms [182].

1.5.3 Summary of Methodological Challenges

Below are listed a few of the key questions regarding future developments in clinically-driven image analysis research:

- How can the validity of non-rigid registration be tested?
- What techniques can be used to favour anatomically reasonable deformations?
- Can meaningful non-rigid correspondences be found between different subjects?
- Can good correspondences be recovered between one subject's scans and their subsequent post-mortem histology?
- Should manual interaction play a large part in non-rigid registration?
- How can unbiased atlases best be generated?
- Can group-wise non-rigid registration be made fast enough for routine use?

- Is it possible to reliably segment structures which have unclear boundaries in MRI?
- How can information from intensities, voxel neighbourhoods, anatomical priors, and manual interaction be most usefully combined for segmentation?

1.6 Statistical Analysis

1.6.1 Introduction

A crucial part of experimental research is the statistical analysis of the results. In many fields the data are typically in the form of one or a few scalar measurements, such as blood-pressure. If there are multiple measurements then their dependence can be investigated with standard methods of multivariate analysis [212]. However, medical images (and various derived data such as segmentations, surfaces, non-rigid registration displacement fields, etc.) typically contain very large numbers of measurements, which in most cases will have a complicated and practically inestimable dependence structure. For example, a typical T1-weighted MR volume of the brain might have $256 \times 256 \times 124$ voxels, in which, firstly, neighbouring voxels will be correlated due to the point-spread function of the acquisition, and, secondly, the underlying anatomy will lead to dependence between both neighbouring and more distant voxels, for example within a particular tissue type.

The multivariate nature of medical images is often ignored, either due to constraints on computational resources, or because limited data make the estimation of covariances unreliable. Investigators may also prefer simpler statistical methods for the ease of communicating their results. Techniques such as VBM [39] treat each voxel independently of the others, allowing simple univariate statistics to be applied.

1.6.2 Basic univariate methods

To start with, a brief description of some simple methods is given, building up to more complex models, before discussing possible avenues for future work.

Student's t-test

The t-test [213] is one of the simplest and most commonly used statistical methods. The single-sample t-test concerns a hypothesis on the mean of a sample; the two-sample t-test considers the difference in means between two groups. The paired t-test is appropriate when the observations in two samples are in correlated pairs, as for example when a single quantity is measured on two occasions, and hence is important in longitudinal studies. The paired t-test is equivalent to a single-sample t-test on the difference between the paired values. The t-value is essentially a measure of 'signal-to-noise'. The numerator is an estimate of the mean or the difference in means between two groups, while the denominator estimates the corresponding standard error. The degrees of freedom, v, typically equal to the number of samples minus the number of parameters estimated from them, can be thought of as characterising the uncertainty in the variance estimate.

ANOVA and F-tests

By considering the increase in residual variance caused by restricting a subset of parameters in a model due to a certain hypothesis (e.g. that some parameters are zero), more general hypotheses can be tested [214]. This is the essential idea behind the F-test and the Analysis of Variance (ANOVA) [213], which can be used to investigate factorial experiments [215], in which *factors* such as group and time-point, take certain *levels* such as control/patient or baseline/repeat. One may be interested in the *main effects* such as differences between patient groups or changes over time, or in the *interactions* between factors, which include the important case of a chronologically-changing group difference — as one would expect with e.g. AD versus control over time.

For longitudinal data, the extension of a paired t-test to ANOVA is known as 'repeated measures' or 'within-subjects' ANOVA [216, 217, 218], and allows one to consider more than two time-points, and to test hypotheses on the interaction between group and time in a principled way. For two groups over two (paired) time-points, a test of the interaction between group and time can be shown to be equivalent to a cross-sectional two-sample t-test of the differences in paired values.

Summary-statistics

For unbalanced longitudinal data, in which measurement times and/or numbers of measurements vary for different subjects, the simplest approach is to perform between-subject analysis of within-subject summary-statistics, for example slopes from regressions against measurement time [219]. The above-mentioned paired-differences are a special case of the slope, with two measurements separated by a unit measurement interval.

This two-stage procedure is sometimes known as the 'NIH method' [220, p.7] and is the most commonly used method for multi-subject fMRI studies, where it is usually called the random-effects approach [221, 222] in contrast to mixed-effects approaches that model both between- and within-subject variability [223, 224], discussed further in 1.6.3.

Randomised Controlled Trials and ANCOVA

A clinically important statistical design is that of the placebo-controlled double-blind 'Randomised Controlled Trial' (RCT) [225]. Baseline measurements are taken, then subjects are randomly allocated to placebo or treatment groups, one or more follow-up measurements are taken after the treatment (e.g. after administering a drug/placebo), and the effects on the groups are analysed. The data are naturally longitudinal, and could be analysed using repeated measures ANOVA or simple summary statistics such as the difference between pre- and post-treatment means.

Frison and Pocock [225] considered an alternative approach, in which the baseline values are used as a covariate for time-averaged follow-up values in a regression model, known as Analysis of Covariance (ANCOVA). The model makes the assumption that the groups shouldn't differ at baseline (due to the nature of the RCT). Frison and Pocock showed that this approach is statistically more powerful than a simple analysis of paired

differences, and also that the latter can be biased in the event of observed baseline group differences, due to 'regression to the mean'.

Vickers has shown that the commonly used summary statistic of percentage change from baseline (i.e. longitudinal difference divided by baseline value) is flawed, and recommends Frison and Pocock's ANCOVA method instead [226]. The same author also suggests that ANOVA in general should be avoided for RCTs, and that regression methods such as ANCOVA should be used in preference [227].

For RCTs in dementia, the interest is usually in whether the *rate* of atrophy and/or cognitive decline is modified by the treatment. Frison and Pocock later [228] extended their method to consider such cases, under the assumption of linearly divergent measurements following treatment. Again, they show that the most obvious summary statistic (the fitted straight-line slope) is less powerful than an ANCOVA-based method.

The General Linear Model

All of the models discussed above, from simple t-tests through to repeated measures ANOVA and ANCOVA, can be implemented in a single general framework. The General Linear Model (LM) [214], describes the vector of dependent variable values in terms of a linear combination of explanatory independent variables or column vectors with additive Gaussian noise. The column vectors may be continuous covariates or categorical variables, e.g. coding variables for parameterising levels of an ANOVA-like experimental design.

The model can be concisely written:

$$y \sim \mathrm{N}(Xb, \sigma^2 I) \tag{1.1}$$

where y is a length n vector, and X is $n \times p$. Appendix A.4 derives a procedure by which any hypothesis that can be written in terms of a linear combination of the coefficients $c^T b = d$ (usually with $d = 0$) can be tested with a t-statistic. This includes many common hypotheses such as 'means differ for groups 1 and 2' or 'time 2 values are higher than time 1'. For more complicated hypotheses, such as 'means differ for all three groups', the null hypothesis involves multiple linear combinations of the parameters, $C^T b = d$, for suitable matrix C and vector d (again usually a zero-vector). An F-statistic can be derived in terms of the increased mean-squared error due to the hypothesis, as shown in the appendix.

Examples of the LM

An unpaired t-test simply models two groups with different means, and tests the hypothesis that the means are equal, it can be implemented as:

$$y^T = [a\ b\ c\ d\ q\ r\ s] \tag{1.2}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \tag{1.3}$$

$$c^T = [1\ \text{-}1] \tag{1.4}$$

where the groups $(\{a, b, c, d\}, \{q, r, s\})$ need not have equal size. The contrast gives the null hypothesis that the first mean minus the second is zero, and hence that they are equal.

For a paired test, each subject has a personal mean, and there is an extra explanatory variable for the effect of time. The null hypothesis is that this extra time effect is zero:

$$y^T = [a_1\ b_1\ c_1\ a_2\ b_2\ c_2] \tag{1.5}$$

$$X^T = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \tag{1.6}$$

$$c^T = [0\ 0\ 0\ 1] \tag{1.7}$$

Unlike simpler expressions for the paired t-test, the above can be extended to more than two time-points. For example, for three times, the design:

$$y^T = [a_1\ b_1\ c_1\ a_2\ b_2\ c_2\ a_3\ b_3\ c_3] \tag{1.8}$$

$$X^T = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \tag{1.9}$$

would allow the testing of the hypothesis that the second time point exceeded[5] the first, $c^T = [0\ 0\ 0\ 1\ 0]$, or that the third exceeded the first, $c^T = [0\ 0\ 0\ 0\ 1]$, or if more appropriate, that the third exceeded the second $c^T = [0\ 0\ 0\ \text{-}1\ 1]$. The F-contrast formed from 'collecting' the first and second (or third) t-contrast vectors into an $m \times 2$ matrix tests the null hypothesis that all time-points are equal.

It is important to note that there are different ways of expressing the same design and hypotheses, with varying ease of interpretation. This becomes particularly relevant if one allows the use of rank-deficient design matrices. Estimable contrasts (see section A.4.5) can be successfully tested in the rank-deficient case, but it may not be immediately obvious

[5] Assuming one-sided tests, as are commonly chosen for t-contrasts in SPM; for a two-sided test the hypothesis is simply that the time-points differ.

what the correct contrast should be if the model is highly over-parametrised. In the examples here, we have preferred to use full-rank designs, derived from consideration of 'extra' effects, rather than introducing covariates for every level of each factor and every combination of levels for interaction terms.

A Comparison of two or more groups over two or more time-points may be achieved with an ANOVA model. For example, with two groups measured over three times, $(\{a, b, c\}, \{q, r\})$ the design could be:

$$y^T = [a_1\ b_1\ c_1\ q_1\ r_1\ a_2\ b_2\ c_2\ q_2\ r_2\ a_3\ b_3\ c_3\ q_3\ r_3] \tag{1.10}$$

$$X^T = \left[\begin{array}{c|c|c} I_5 & I_5 & I_5 \\ \hline zeros_{1,5} & ones_{1,5} & zeros_{1,5} \\ \hline zeros_{1,5} & zeros_{1,5} & ones_{1,5} \\ \hline zeros_{1,5} & [0\ 0\ 0\ 1\ 1] & zeros_{1,5} \\ \hline zeros_{1,5} & zeros_{1,5} & [0\ 0\ 0\ 1\ 1] \end{array}\right] \tag{1.11}$$

where I_n is the $n \times n$ identity matrix, $zeros_{n,m}$ is an $n \times m$ matrix of zeros (and similarly for ones).

The final two columns give the interactions between group and time, or, in other words, the additional time effect for the second group. The contrast $c^T = [0\ 0\ 0\ 0\ 0\ 0$ -1 $1]$ would test whether the difference between the groups was growing over time. This kind of analysis can become rather complicated, and, as mentioned earlier, some authors recommend summary statistics instead [225, 227, 228], both for reasons of simplicity and of statistical superiority under certain experimental designs.

The ANCOVA method [225] for the analysis of a simple two-group RCT with two time-points can be implemented by modelling the follow-up measurements as the dependent variable, and putting the baselines into the design matrix as a covariate, as follows:

$$y^T = [a_2\ b_2\ c_2\ q_2\ r_2] \tag{1.12}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ a_1 & b_1 & c_1 & q_1 & r_1 \end{bmatrix} \tag{1.13}$$

$$c^T = [-1\ 1\ 0] \tag{1.14}$$

where the contrast will directly test the drug effect.

1.6.3 Extensions to the linear model

Appendix A.4 presents the multivariate general linear model in some detail. Here, we briefly mention some of the other extensions.

While quite broadly applicable, the LM framework is based on assumptions that may be overly restrictive in certain situations. In equation (1.1) the mean is modelled with a linear combination of terms, and the noise (and hence likelihood, in this linear model) is assumed Gaussian.

Generalized Linear Models [229] extend the framework to include other noise distributions, such as Poisson, and allow for a non-identity 'link-function' that models some non-linear function of the mean as a linear combination. These extensions allow the models to be used for applications such as logistic regression, or the analysis of discrete variables such as count data or binary data. The price of this increased flexibility is that the simple closed form solutions must be replaced with iterative estimation methods.

The LM also assumes a very basic structure for the residual error. In equation (1.1) the covariance matrix is a simple scaled identity matrix, meaning that the error values are independent and identically distributed. This may break down in several ways, such as heterogeneous variance in different groups or at different times, or correlated residuals. The latter can be a serious problem in longitudinal studies, since even with a repeated-measures model there may be remaining correlations over time in the error.

It is simple to derive expressions similar to those in section 1.6.2 for a completely general covariance matrix Σ, or, equivalently, the data can be 'pre-whitened' by multiplying y and X by $\Sigma^{-1/2}$. However, this requires that the entire $n \times n$ covariance matrix be known a priori or accurately estimable from the available data — something which is rarely true in traditional analyses. However, neuroimaging presents an interesting special case: if one is willing assume a common covariance structure over voxels (scaled by a non-stationary variance), then it becomes possible to pool all the voxels,[6] permitting estimation of the variance components with effectively total precision, meaning they can subsequently be treated as known, as advocated by Glaser and Friston [230].

Multilevel or hierarchical models [231], also known as mixed effects models [223, 229], aim to allow more general correlation structures than the simple LM without having to specify or estimate all possible variances and covariances between variables, and without having to make assumptions about the constancy of variance components over voxels. They achieve this by breaking down a complete model into a hierarchy of levels with simpler but interacting correlation structures, or equivalently by introducing 'random effects' to model the variance components in addition to the 'fixed effects' linear model of the mean (leading to an overall 'mixed effects' model). The models have recently been described in a general 'latent variable' framework [232] which encompasses several other important statistical techniques such as structural equation models.

1.6.4 The Multiple Testing Problem

Conventional hypothesis testing is based on the idea of controlling the false-positive rate or type I error, α. If multiple statistical tests are performed then the overall probability of getting false-positives clearly increases. This problem originally arose in situations with relatively small degrees of multiplicity, such as ANOVA designs in which several comparisons of particular levels' means may be of interest, in addition to the main effect hypothesis of all levels' means being equal. E.g. in a one-way ANOVA with four levels A, B, C and D, the investigator could test ${}_4C_2 = 6$ pairs of means. Multiple comparison

[6] To be precise, the SPM software pools only those voxels which exceed a pre-specified main-effect threshold.

procedures are available that aim to control the chance of false positives if several such tests are performed. Several such procedures have been developed [233]; we briefly discuss two simple examples below, before considering other approaches more suited to the large number of multiple tests performed in imaging studies. First, it is necessary to define more precisely the basic concepts.

With multiple comparisons, several different measures of false-positive rate are available [234]. The family-wise error rate (FWE) is the chance of any false positives occuring in any of the individual tests. When multiple comparisons are being made, it is possible that a mixture of null and alternative hypotheses could be true, it is therefore necessary to distinguish two senses in which FWE can be controlled. 'Weak' control of FWE means the chance of any false positives is no greater than the nominal level given that the null hypothesis is true for all of the comparisons (a 'complete' null hypothesis). 'Strong' control applies to the mixed or partial null case, and requires that the chance of any false positives be controlled over any subset of comparisons for which the null hypothesis holds.

Fisher's Least Significant Difference (LSD), sometimes known as Fisher's Protected LSD method, follows the logic that we can be 'protected' from errors in the individual comparisons, by only performing them if a main effects test over all comparisons indicates that there is a significant difference. This approach provides weak control of FWE for the simple reason that the main effects test controls false positives at the nominal level if there is no true effect in any of the comparisons. However, LSD does not offer strong control of FWE. One or more true alternative hypotheses can be sufficient for the main effects test to be rejected, allowing all the protected individual tests to be performed; but if several comparisons are made for which the null hypothesis is true, the probability of one or more false positives occuring in those tests may rise above the nominal level. The important point to note from this is that one can no longer be confident that individual rejected tests are significant.

Bonferroni correction provides strong control over false positives, allowing rejection of individual hypotheses. The method relies on a conservative inequality regarding the probability of any false-positive. If F_i denotes the event of falsely rejecting the i^{th} hypothesis, we have $\Pr(F_1 \cup F_2) = \Pr(F_1) + \Pr(F_2) - \Pr(F_1 \cap F_2) \leq \Pr(F_1) + \Pr(F_2)$, regardless of the dependence of F_1 and F_2 or of the presence of other tests with true or false null hypotheses. Note that the greater the dependency, the more conservative the inequality is. More generally, Boole's inequality holds:

$$\Pr\left(\bigcup_i F_i\right) \leq \sum_i \Pr\left(F_i\right). \tag{1.15}$$

If one can ensure that the chances of individual false positives are all less than or equal to α_B, then the right hand side of the inequality is no greater than $N\alpha_B$ for N tests, therefore if the individual hypotheses are tested at a level $\alpha_B = \alpha_0/N$ FWE will be controlled at the nominal level α_0. The thresholds for the individual hypotheses could be different [234],[7] but it is common to consider a single threshold. For example a critical t-value

[7] An example of different thresholds occurs with the use of step-down permutation-based correction of

corresponding to a p-value of α_0/N would control FWE at α_0 over a set of N t-statistics. One can equivalently consider the Bonferroni-corrected p-values of N tests to be given by N times the uncorrected p-values.

Multiple comparison correction in ANOVA-like scenarios has been the subject of some controversy [235, 236]. One argument is that if each of several possible tests are of independent scientific interest, then they should not be disadvantaged in terms of power simply because they happen to have been studied simultaneously. In situations with relatively small numbers of multiple comparisons, the most reasonable thing to do may be to note that multiple tests have been carried out (even if only the significant ones are discussed), letting the readers judge for themselves to what extent the significance of the reported findings should be lowered.

Volumetric MR images usually contain a very large number of voxels (of the order of 10^5–10^7), if mass-univariate voxel-wise statistical tests are carried out for each of these voxels, the scale of the resultant multiple testing problem[8] is very different to the above examples. Different voxels are typically not of independent scientific interest, and it would be very difficult for the reader to accurately judge the extent of the multiplicity, which varies with the level of dependence between voxels. It is clear therefore that some form of correction for multiple testing is necessary. There is also usually a desire to be able to localise results to particular brain regions by declaring certain voxels to be significant, possibly in the presence of other voxels for which the alternative hypothesis is true, which requires strong control of family-wise error. However, use of the Bonferroni correction with such large numbers of dependent tests would be extremely conservative, resulting in a very low sensitivity to true-positive results.

Medical images typically have a degree of local correlation, due to the nature of the biological structure and the regional nature of pathology as well as the point-spread function of the acquisition process. In the popular technique of Statistical Parametric Mapping [238] the images undergo substantial further spatial smoothing, guaranteeing that neighbouring voxels are highly statistically dependent. These correlated tests should clearly not be corrected for as though they were all independent. An intuitively appealing idea is to estimate the correlation in some way, and from this, the effective number of independent tests; one could then use Bonferroni correction with this smaller number in place of the total number of voxels. However, this turns out not to be a succesful approach for particularly smooth neuroimaging data, as demonstrated in [234].[9] Instead, methods have been derived which can control FWE more accurately than Bonferroni by considering the distribution of the maximum of the multiple test statistics.

In order for one or more tests to be rejected at a certain threshold, it is necessary for

FWE, presented later in section 2.3.1.

[8] We use the term 'multiple testing' here instead of 'multiple comparisons', following [237], where the latter term is recommended for multiple comparisons of means in ANOVA models, and the former term is favoured for more general multiple testing.

[9] In particular, note that using the number of resolution elements (resels) [239] as the number of independent tests in a Bonferroni correction fails to control FWE [234].

the maximum of the statistics to be above this threshold, i.e.

$$\Pr\left(\bigcup_i \{T_i \geq T_c\} \,\middle|\, H_0\right) = \Pr(\max_i T_i \geq T_c | H_0)$$

where T_i are the test statistics, and T_c is a critical statistic threshold. If T_c is chosen as the $100(1-\alpha_0)$ percentile of the distribution of the maximum statistic under a complete null hypothesis then the equality implies that FWE will be weakly controlled. Under an assumption known as subset pivotality [240], which means that the null distribution of a subset of tests is independent of the truth of other null hypotheses, T_c will in fact strongly control FWE [234]. Mass-univariate imaging statistics satisfy subset pivotality because there are no constraints between the the null hypotheses of different voxels [234].

However, the maximum distribution is typically not known, or not available in closed form. If the multiple tests are independent, then the maximum distribution is the product of the individual cumulative distribution functions, but if there is a complex dependence structure between the tests (as will likely be the case between voxels in imaging studies), the maximum distribution must be approximated somehow. Section 2.3 explains how permutation testing can be used to derive an empirical estimate of the null distribution, and of the maximum distribution in multiple testing. Here, we briefly discuss a very popular parametric approximation to the maximum distribution.

Random Field Theory (RFT) [241] can be applied to two- or three-dimensional continuous spatial fields of statistics. If discretely sampled images of statistics are sufficiently finely sampled in relation to their smoothness[10] then RFT results should be approximately valid. For a particular threshold T_c the excursion set of a random field defined over Ω is $\{s \in \Omega : T(s) > T_c\}$. The Euler Characteristic of the excursion set is a topological measure which, at sufficiently high thresholds T_c approximately counts the number of suprathreshold clusters [234]. If the threshold is high enough for the probability of multiple clusters to be negligible, the expected value of the Euler Characteristic is approximately equal to its probability of being non-zero. The relevance of this is that the probability of having any suprathreshold clusters is the FWE, and the expected value of the Euler Characteristic is an approximation to this which can in turn be approximated in closed form for a number of statistical fields including Gaussian, t, χ^2, F, and T^2 [242, 243].[11] This provides parametric expressions for FWE corrected thresholds or p-values, if the necessary assumptions for RFT hold [234]. It is also possible to use RFT to derive FWE-corrected p-values of clusters of contiguous voxels above a prespecified threshold (e.g. an uncorrected p-value of 0.001) based on their size or mass (the integral of suprathreshold intensities) [245, 246].

Power and sample-size calculations are more challenging when correcting for multiple-tests, but a recent paper has presented results for voxel-level FWE calculations based on the theory of non-central random fields [247].

[10] A rule of thumb of three voxels full-width at half maximum smoothness is mentioned and evaluated in [234].

[11] An interesting point here is that the standard single-statistic transformations from e.g. t to standard normal Z or correlation coefficient ρ do not correctly transform between t and Z or t and ρ random fields [234, 244].

In some situations, even relatively accurate control of FWE can still be undesirably conservative, in the sense that the large number of false negatives or type II errors might be of more concern to the investigator than the exact number of false positives. This motivated the application of False Discovery Rate (FDR) correction [248, 249] to neuroimaging [250]. FDR aims to correct for the proportion of false-positives among the rejected null hypotheses, rather than the probability of falsely rejecting a single null hypothesis. The two approaches are equivalent if all null hypotheses are true (i.e. there are no significant voxels anywhere), which means that FDR provies weak control of FWE. However, FDR is less conservative in the (often quite likely) event that there are some truly significant voxels somewhere in a particular test. FDR may also be applied to uncorrected RFT p-values for cluster-size [251].

1.7 Conclusion

This chapter has presented the clinical background for the application, and introduced the basic image processing methods of registration and segmentation that are essential to later work. The technique of voxel-based morphometry has been introduced, which is returned to in detail in chapter 3. Basic statistical methods have been outlined, with a focus on the multiple comparison problem, so that later discussion of the relative merits of different types of correction procedure can be appreciated. The next chapter studies in great detail a particular class of statistical method, developing a permutation-testing framework that is then used in chapter 4, which attempts to improve upon VBM by incorporating certain multivariate aspects of the data.

Bibliography

[1] "Website of the Mayo Clinic." [Online]. Available: http://www.mayoclinic.com/health/dementia ˆ13

[2] "Website of the National Institute of Neurological Disorders and Stroke." [Online]. Available: http://www.ninds.nih.gov/disorders/dementias/dementia.htm ˆ13

[3] "Fact Sheets from Alzheimer's Association." [Online]. Available: http://www.alz.org/alzheimers_disease_publications.asp ˆ13, 14

[4] D. A. Evans, H. H. Funkenstein, M. S. Albert, P. A. Scherr, N. R. Cook, M. J. Chown, L. E. Hebert, C. H. Hennekens, and J. O. Taylor, "Prevalence of Alzheimer's disease in a community population of older persons. Higher than previously reported." *JAMA*, vol. 262, no. 18, pp. 2551–2556, Nov. 1989. ˆ13

[5] C. P. Ferri, M. Prince, C. Brayne, H. Brodaty, L. Fratiglioni, M. Ganguli, K. Hall, K. Hasegawa, H. Hendrie, Y. Huang, A. Jorm, C. Mathers, P. R. Menezes, E. Rimmer, M. Scazufca, and A. D. International, "Global prevalence of dementia: a Delphi consensus study." *Lancet*, vol. 366, no. 9503, pp. 2112–2117, Dec. 2005. ˆ13

[6] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""Mini-mental state". A practical method for grading the cognitive state of patients for the clinician." *J Psychiatr Res*, vol. 12, no. 3, pp. 189–198, Nov. 1975. ˆ14

[7] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, "Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease." *Neurology*, vol. 34, no. 7, pp. 939–944, Jul. 1984. ˆ14

[8] N. C. Fox, E. K. Warrington, P. A. Freeborough, P. Hartikainen, A. M. Kennedy, J. M. Stevens, and M. N. Rossor, "Presymptomatic hippocampal atrophy in Alzheimer's disease. A longitudinal MRI study." *Brain*, vol. 119 (Pt 6), pp. 2001–2007, Dec. 1996. ˆ14

[9] D. Head, A. Z. Snyder, L. E. Girton, J. C. Morris, and R. L. Buckner, "Frontal-hippocampal double dissociation between normal aging and Alzheimer's disease." *Cereb Cortex*, vol. 15, no. 6, pp. 732–739, Jun. 2005. ˆ14

[10] N. C. Fox, P. A. Freeborough, and M. N. Rossor, "Visualisation and quantification of rates of atrophy in Alzheimer's disease." *Lancet*, vol. 348, no. 9020, pp. 94–97, Jul. 1996. ˆ14

[11] N. C. Fox, W. R. Crum, R. I. Scahill, J. M. Stevens, J. C. Janssen, and M. N. Rossor, "Imaging of onset and progression of Alzheimer's disease with voxel-compression mapping of serial magnetic resonance images." *Lancet*, vol. 358, no. 9277, pp. 201–205, Jul. 2001. ˆ14, 16

[12] K. Kantarci and C. R. Jack, "Neuroimaging in Alzheimer disease: an evidence-based review." *Neuroimaging Clin N Am*, vol. 13, no. 2, pp. 197–209, May 2003. ˆ14, 20

[13] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen, "Mild cognitive impairment: clinical characterization and outcome." *Arch Neurol*, vol. 56, no. 3, pp. 303–308, Mar. 1999. ˆ14

[14] J. C. Morris, M. Storandt, J. P. Miller, D. W. McKeel, J. L. Price, E. H. Rubin, and L. Berg, "Mild cognitive impairment represents early-stage Alzheimer disease." *Arch Neurol*, vol. 58, no. 3, pp. 397–405, Mar. 2001. ˆ14

[15] G. B. Karas, P. Scheltens, S. A. R. B. Rombouts, P. J. Visser, R. A. van Schijndel, N. C. Fox, and F. Barkhof, "Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease." *Neuroimage*, vol. 23, no. 2, pp. 708–716, Oct. 2004. ˆ14, 16

[16] M. N. Rossor, N. C. Fox, P. A. Freeborough, and R. J. Harvey, "Clinical features of sporadic and familial Alzheimer's disease." *Neurodegeneration*, vol. 5, no. 4, pp. 393–397, Dec. 1996. ˆ14

[17] H. Braak and E. Braak, "Neuropathological stageing of Alzheimer-related changes." *Acta Neuropathol (Berl)*, vol. 82, no. 4, pp. 239–259, 1991. ˆ14

[18] N. C. Fox and M. N. Rossor, "Seeing what Alzheimer saw - with magnetic resonance microscopy." *Nat Med*, vol. 6, no. 1, pp. 20–21, Jan. 2000. ˆ14

[19] H. Benveniste, G. Einstein, K. R. Kim, C. Hulette, and G. A. Johnson, "Detection of neuritic plaques in Alzheimer's disease by magnetic resonance microscopy." *Proc Natl Acad Sci U S A*, vol. 96, no. 24, pp. 14 079–14 084, Nov. 1999. ˆ14

[20] C. R. Jack, M. Garwood, T. M. Wengenack, B. Borowski, G. L. Curran, J. Lin, G. Adriany, O. H. J. Gröhn, R. Grimm, and J. F. Poduslo, "In vivo visualization of Alzheimer's amyloid plaques by magnetic resonance imaging in transgenic mice without a contrast agent." *Magn Reson Med*, vol. 52, no. 6, pp. 1263–1271, Dec. 2004. ˆ14

[21] M. Higuchi, N. Iwata, Y. Matsuba, K. Sato, K. Sasamoto, and T. C. Saido, "19F and 1H MRI detection of amyloid beta plaques in vivo." *Nat Neurosci*, vol. 8, no. 4, pp. 527–533, Apr. 2005. ˆ14

[22] W. E. Klunk, H. Engler, A. Nordberg, Y. Wang, G. Blomqvist, D. P. Holt, M. Bergström, I. Savitcheva, G. feng Huang, S. Estrada, B. Ausén, M. L. Debnath, J. Barletta, J. C. Price, J. Sandell, B. J. Lopresti, A. Wall, P. Koivisto, G. Antoni, C. A. Mathis, and B. Långström, "Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B." *Ann Neurol*, vol. 55, no. 3, pp. 306–319, Mar. 2004. ˆ14

[23] A. M. Fagan, M. A. Mintun, R. H. Mach, S.-Y. Lee, C. S. Dence, A. R. Shah, G. N. LaRossa, M. L. Spinner, W. E. Klunk, C. A. Mathis, S. T. DeKosky, J. C. Morris, and D. M. Holtzman, "Inverse relation between in vivo amyloid imaging load and cerebrospinal fluid Abeta42 in humans." *Ann Neurol*, vol. 59, no. 3, pp. 512–519, Mar. 2006. ˆ14

[24] M. Hintersteiner, A. Enz, P. Frey, A.-L. Jaton, W. Kinzy, R. Kneuer, U. Neumann, M. Rudin, M. Staufenbiel, M. Stoeckli, K.-H. Wiederhold, and H.-U. Gremlich, "In vivo detection of amyloid-beta deposits by near-infrared imaging using an oxazine-derivative probe." *Nat Biotechnol*, vol. 23, no. 5, pp. 577–583, May 2005. ˆ14

[25] M. J. de Leon, S. DeSanti, R. Zinkowski, P. D. Mehta, D. Pratico, S. Segal, C. Clark, D. Kerkman, J. DeBernardis, J. Li, L. Lair, B. Reisberg, W. Tsui, and H. Rusinek, "MRI and CSF studies in the early diagnosis of Alzheimer's disease." *J Intern Med*, vol. 256, no. 3, pp. 205–223, Sep. 2004. ˆ14

[26] D. J. Selkoe, "Aging, amyloid, and Alzheimer's disease: a perspective in honor of Carl Cotman." *Neurochem Res*, vol. 28, no. 11, pp. 1705–1713, Nov. 2003. ˆ14

[27] V. T. Marchesi, "An alternative interpretation of the amyloid Abeta hypothesis with regard to the pathogenesis of Alzheimer's disease." *Proc Natl Acad Sci U S A*, vol. 102, no. 26, pp. 9093–9098, Jun. 2005. ˆ15

[28] H. F. Dovey, V. John, J. P. Anderson, L. Z. Chen, P. de Saint Andrieu, L. Y. Fang, S. B. Freedman, B. Folmer, E. Goldbach, E. J. Holsztynska, K. L. Hu, K. L. Johnson-Wood, S. L. Kennedy, D. Kholodenko, J. E. Knops, L. H. Latimer, M. Lee, Z. Liao, I. M. Lieberburg, R. N. Motter, L. C. Mutter, J. Nietz, K. P. Quinn, K. L. Sacchi, P. A. Seubert, G. M. Shopp, E. D. Thorsett, J. S. Tung, J. Wu, S. Yang, C. T. Yin, D. B. Schenk, P. C. May, L. D. Altstiel, M. H. Bender, L. N. Boggs, T. C. Britton, J. C. Clemens, D. L. Czilli, D. K. Dieckman-McGinty, J. J. Droste, K. S. Fuson, B. D. Gitter, P. A. Hyslop, E. M. Johnstone, W. Y. Li, S. P. Little, T. E. Mabry, F. D. Miller, and J. E. Audia, "Functional gamma-secretase inhibitors reduce beta-amyloid peptide levels in brain." *J Neurochem*, vol. 76, no. 1, pp. 173–181, Jan. 2001. ˆ15

[29] S. Gandy, "The role of cerebral amyloid beta accumulation in common forms of Alzheimer disease." *J Clin Invest*, vol. 115, no. 5, pp. 1121–1129, May 2005. ˆ15

[30] D. Schenk, R. Barbour, W. Dunn, G. Gordon, H. Grajeda, T. Guido, K. Hu, J. Huang, K. Johnson-Wood, K. Khan, D. Kholodenko, M. Lee, Z. Liao, I. Lieberburg, R. Motter, L. Mutter, F. Soriano, G. Shopp, N. Vasquez, C. Vandevert, S. Walker, M. Wogulis, T. Yednock, D. Games, and P. Seubert, "Immunization with amyloid-beta attenuates Alzheimer-disease-like pathology in the PDAPP mouse." *Nature*, vol. 400, no. 6740, pp. 173–177, Jul. 1999. ˆ15

[31] S. Gilman, M. Koller, R. S. Black, L. Jenkins, S. G. Griffith, N. C. Fox, L. Eisner, L. Kirby, M. B. Rovira, F. Forette, J.-M. Orgogozo, and A. N. S.-.-. S. Team, "Clinical effects of Abeta immunization (AN1792) in patients with AD in an interrupted trial." *Neurology*, vol. 64, no. 9, pp. 1553–1562, May 2005. ˆ15

[32] E. Masliah, L. Hansen, A. Adame, L. Crews, F. Bard, C. Lee, P. Seubert, D. Games, L. Kirby, and D. Schenk, "Abeta vaccination effects on plaque pathology in the absence of encephalitis in Alzheimer disease." *Neurology*, vol. 64, no. 1, pp. 129–131, Jan. 2005. ˆ15

[33] N. C. Fox, R. S. Black, S. Gilman, M. N. Rossor, S. G. Griffith, L. Jenkins, and M. Koller, "Effects of A-beta immunization (AN1792) on MRI measures of cerebral volume in Alzheimer disease." *Neurology*, vol. 64, no. 9, pp. 1563–1572, May 2005, for the AN1792(QS-21)-201 Study Team. ˆ15

[34] P. A. Freeborough, N. C. Fox, and R. I. Kitney, "Interactive algorithms for the segmentation and quantitation of 3-D MRI brain scans." *Comput Methods Programs Biomed*, vol. 53, no. 1, pp. 15–25, May 1997. ˆ16

[35] F. L. Bookstein, "Linear methods for nonlinear maps: Procrustes fits, Thin-Plate Splines, and the biometric analysis of shape variability," in *Brain Warping*, A. W. Toga, Ed. Academic Press, 1999, ch. 10, pp. 157–181. ˆ16, 24

[36] P. A. Freeborough and N. C. Fox, "The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI." *IEEE Trans. Med. Imag.*, vol. 16, no. 5, pp. 623–629, Oct. 1997. ˆ16

[37] S. M. Smith, Y. Zhang, M. Jenkinson, J. Chen, P. M. Matthews, A. Federico, and N. D. Stefano, "Accurate, robust, and automated longitudinal and cross-sectional brain change analysis." *Neuroimage*, vol. 17, no. 1, pp. 479–489, Sep. 2002. ˆ16

[38] I. C. Wright, P. K. McGuire, J. B. Poline, J. M. Travere, R. M. Murray, C. D. Frith, R. S. Frackowiak, and K. J. Friston, "A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia." *Neuroimage*, vol. 2, no. 4, pp. 244–252, Dec. 1995. ˆ16

[39] J. Ashburner and K. J. Friston, "Voxel-based morphometry–the methods." *Neuroimage*, vol. 11, no. 6 Pt 1, pp. 805–821, Jun. 2000. ˆ16, 17, 28

[40] ——, "Why voxel-based morphometry should be used." *Neuroimage*, vol. 14, no. 6, pp. 1238–1243, Dec. 2001. ˆ16

[41] C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. Henson, K. J. Friston, and R. S. Frackowiak, "A voxel-based morphometric study of ageing in 465 normal adult human brains." *Neuroimage*, vol. 14, no. 1 Pt 1, pp. 21–36, Jul. 2001. ˆ16

[42] A. Mechelli, C. J. Price, K. J. Friston, and J. Ashburner, "Voxel-based morphometry of the human brain: Methods and applications," *Current Medical Imaging Reviews*, vol. 1, no. 1, pp. 1–9, 2005. ˆ16

[43] J. C. Baron, G. Chételat, B. Desgranges, G. Perchey, B. Landeau, V. de la Sayette, and F. Eustache, "In vivo mapping of gray matter loss with voxel-based morphometry in mild Alzheimer's disease." *Neuroimage*, vol. 14, no. 2, pp. 298–309, Aug. 2001. ˆ16

[44] G. B. Karas, E. J. Burton, S. A. R. B. Rombouts, R. A. van Schijndel, J. T. O'Brien, P. Scheltens, I. G. McKeith, D. Williams, C. Ballard, and F. Barkhof, "A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry." *Neuroimage*, vol. 18, no. 4, pp. 895–907, Apr. 2003. ˆ16

[45] F. L. Bookstein, ""Voxel-based morphometry" should not be used with imperfectly registered images." *Neuroimage*, vol. 14, no. 6, pp. 1454–1462, Dec. 2001. ˆ16

[46] C. Davatzikos, "Why voxel-based morphometric analysis should be used with great caution when characterizing group differences." *Neuroimage*, vol. 23, no. 1, pp. 17–20, Sep. 2004. ˆ16, 17

[47] A. Bartsch, N. Bendszus, N. D. Stefano, G. Homola, and S. Smith, "Extending SIENA for a multi-subject statistical analysis of sample-specific cerebral edge shifts: Substantiation of early brain regeneration through abstinence from alcoholism." in *Tenth Int. Conf. on Functional Mapping of the Human Brain*, 2004. ˆ16

[48] R. I. Scahill, J. M. Schott, J. M. Stevens, M. N. Rossor, and N. C. Fox, "Mapping the evolution of regional atrophy in Alzheimer's disease: unbiased analysis of fluid-registered serial MRI." *Proc Natl Acad Sci USA*, vol. 99, no. 7, pp. 4703–4707, Apr. 2002. ˆ17

[49] C. Studholme, V. Cardenas, R. Blumenfeld, N. Schuff, H. J. Rosen, B. Miller, and M. Weiner, "Deformation tensor morphometry of semantic dementia with quantitative validation." *Neuroimage*, vol. 21, no. 4, pp. 1387–1398, Apr. 2004. ˆ17

[50] G. Chételat, B. Landeau, F. Eustache, F. Mézenge, F. Viader, V. de la Sayette, B. Desgranges, and J.-C. Baron, "Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study." *Neuroimage*, vol. 27, no. 4, pp. 934–946, Oct. 2005. ˆ17

[51] P. M. Thompson, D. MacDonald, M. S. Mega, C. J. Holmes, A. C. Evans, and A. W. Toga, "Detection and mapping of abnormal brain structure with a probabilistic atlas of cortical surfaces." *J Comput Assist Tomogr*, vol. 21, no. 4, pp. 567–581, 1997. ˆ17

[52] A. M. Dale, B. Fischl, and M. I. Sereno, "Cortical surface-based analysis. I. Segmentation and surface reconstruction." *Neuroimage*, vol. 9, no. 2, pp. 179–194, Feb. 1999. ˆ17

[53] B. Fischl, M. I. Sereno, and A. M. Dale, "Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system." *Neuroimage*, vol. 9, no. 2, pp. 195–207, Feb. 1999. ˆ17

[54] B. Fischl and A. M. Dale, "Measuring the thickness of the human cerebral cortex from magnetic resonance images." *Proc Natl Acad Sci U S A*, vol. 97, no. 20, pp. 11 050–11 055, Sep. 2000. ˆ17

[55] D. W. Shattuck and R. M. Leahy, "BrainSuite: an automated cortical surface identification tool." *Med Image Anal*, vol. 6, no. 2, pp. 129–142, Jun. 2002. ˆ17

[56] T. Liu, D. Shen, and C. Davatzikos, "Deformable registration of cortical structures via hybrid volumetric and surface warping." *Neuroimage*, vol. 22, no. 4, pp. 1790–1801, Aug. 2004. ˆ17, 25

[57] P. M. Thompson, K. M. Hayashi, E. R. Sowell, N. Gogtay, J. N. Giedd, J. L. Rapoport, G. I. de Zubicaray, A. L. Janke, S. E. Rose, J. Semple, D. M. Doddrell, Y. Wang, T. G. M. van Erp, T. D. Cannon, and A. W. Toga, "Mapping cortical change in Alzheimer's disease, brain development, and schizophrenia." *Neuroimage*, vol. 23 Suppl 1, pp. S2–18, 2004. ˆ17

[58] P. M. Thompson, K. M. Hayashi, G. de Zubicaray, A. L. Janke, S. E. Rose, J. Semple, D. Herman, M. S. Hong, S. S. Dittmer, D. M. Doddrell, and A. W. Toga, "Dynamics of gray matter loss in Alzheimer's disease." *J Neurosci*, vol. 23, no. 3, pp. 994–1005, Feb. 2003. ˆ17

[59] D. Shen, S. Moffat, S. M. Resnick, and C. Davatzikos, "Measuring size and shape of the hippocampus in MR images using a deformable shape model." *Neuroimage*, vol. 15, no. 2, pp. 422–434, Feb. 2002. ˆ17, 26

[60] D. McRobbie, E. Moore, M. Graves, and M. Prince, *MRI from Picture to Proton*. Cambridge University Press, 2003. ˆ17

[61] Z.-P. Liang and P. C. Lauterbur, *Principles of Magnetic Resonance Imaging: a Signal Processing Perspective*. IEEE Press / Wiley, 2000. ˆ17

[62] E. Haacke, R. Brown, M. Thompson, and R. Venkatesan, *Magnetic resonance imaging: physical principles and sequence design*. Wiley, 1999. ˆ17

[63] T. Sumanaweera, G. Glover, T. Binford, and J. Adler, "MR susceptibility misregistration correction," *IEEE Trans. Med. Imag.*, vol. 12, no. 2, pp. 251–259, 1993. ˆ18, 19

[64] M. O'Donnell and W. A. Edelstein, "NMR imaging in the presence of magnetic field inhomogeneities and gradient field nonlinearities." *Med Phys*, vol. 12, no. 1, pp. 20–26, 1985. ˆ18

[65] K. Sekihara, M. Kuroda, and H. Kohno, "Image restoration from non-uniform magnetic field influence for direct Fourier NMR imaging." *Phys Med Biol*, vol. 29, no. 1, pp. 15–24, Jan. 1984. ˆ18

[66] A. Janke, H. Zhao, G. J. Cowin, G. J. Galloway, and D. M. Doddrell, "Use of spherical harmonic deconvolution methods to compensate for nonlinear gradient effects on MRI images." *Magn Reson Med*, vol. 52, no. 1, pp. 115–122, Jul. 2004. ˆ18

[67] G. Glover and N. Pelc, "Method for correcting image distortion due to gradient nonuniformity," US patentus 4 591 789, 1986. [Online]. Available: http://www.google.com/patents?id=5-M4AAAAEBAJ ˆ18

[68] S. J. Doran, L. Charles-Edwards, S. A. Reinsberg, and M. O. Leach, "A complete distortion correction for MR images: I. Gradient warp correction." *Phys Med Biol*, vol. 50, no. 7, pp. 1343–1361, Apr. 2005. ˆ18

[69] L. Schad, S. Lott, F. Schmitt, V. Sturm, and W. J. Lorenz, "Correction of spatial distortion in MR imaging: a prerequisite for accurate stereotaxy." *J Comput Assist Tomogr*, vol. 11, no. 3, pp. 499–505, 1987. ˆ19

[70] S. Langlois, M. Desvignes, J. M. Constans, and M. Revenu, "MRI geometric distortion: a simple approach to correcting the effects of non-linear gradient fields." *J Magn Reson Imaging*, vol. 9, no. 6, pp. 821–831, Jun. 1999. ˆ19

[71] M. M. Breeuwer, M. Holden, and W. Zylka, "Detection and correction of geometric distortion in 3D MR images," in *Proceedings of SPIE Medical Imaging*, vol. 4322, 2001, pp. 1110–1120. ˆ19

[72] M. Holden, M. M. Breeuwer, K. McLeish, D. J. Hawkes, S. F. Keevil, and D. L. G. Hill, "Sources and correction of higher-order geometrical distortion for serial MR brain imaging," in *Proceedings of SPIE Medical Imaging*, vol. 4322, 2001. ˆ19

[73] A. Littmann, J. Guehring, C. Buechel, and H.-S. Stiehl, "Acquisition-related morphological variability in structural MRI." *Acad Radiol*, vol. 13, no. 9, pp. 1055–1061, Sep. 2006. ˆ19

[74] D. Wang, D. M. Doddrell, and G. Cowin, "A novel phantom and method for comprehensive 3-dimensional measurement and correction of geometric distortion in magnetic resonance imaging." *Magn Reson Imaging*, vol. 22, no. 4, pp. 529–542, May 2004. ˆ19

[75] D. Wang, W. Strugnell, G. Cowin, D. M. Doddrell, and R. Slaughter, "Geometric distortion in clinical MRI systems Part II: correction using a 3D phantom." *Magn Reson Imaging*, vol. 22, no. 9, pp. 1223–1232, Nov. 2004. ˆ19

[76] ——, "Geometric distortion in clinical MRI systems Part I: evaluation using a 3D phantom." *Magn Reson Imaging*, vol. 22, no. 9, pp. 1211–1221, Nov. 2004. ˆ19

[77] D. L. Hill, C. R. Maurer, C. Studholme, J. M. Fitzpatrick, and D. J. Hawkes, "Correcting scaling errors in tomographic images using a nine degree of freedom registration algorithm." *J Comput Assist Tomogr*, vol. 22, no. 2, pp. 317–323, 1998. ˆ19

[78] C. Yu, M. L. Apuzzo, C. S. Zee, and Z. Petrovich, "A phantom study of the geometric accuracy of computed tomographic and magnetic resonance imaging stereotactic localization with the Leksell stereotactic system." *Neurosurgery*, vol. 48, no. 5, pp. 1092–8; discussion 1098–9, May 2001. ˆ19

[79] H. Chang and J. Fitzpatrick, "A technique for accurate magnetic resonance imaging in the presence of field inhomogeneities," *IEEE Trans. Med. Imag.*, vol. 11, no. 3, pp. 319–329, Sep. 1992. ˆ19

[80] H. Chang and J. M. Fitzpatrick, "Geometrical image transformation to compensate for MRI distortions," in *Proc. SPIE Medical Imaging IV: Image Processing*, vol. 1233, Jul. 1990, pp. 116–127. [Online]. Available: http://link.aip.org/link/?PSISDG/1233/116/1 ˆ19

[81] S. A. Kannengiesser, Y. Wang, and E. M. Haacke, "Geometric distortion correction in gradient-echo imaging by use of dynamic time warping." *Magn Reson Med*, vol. 42, no. 3, pp. 585–590, Sep. 1999. ˆ19

[82] S. A. Reinsberg, S. J. Doran, E. M. Charles-Edwards, and M. O. Leach, "A complete distortion correction for MR images: II. Rectification of static-field inhomogeneities by similarity-based profile mapping." *Phys Med Biol*, vol. 50, no. 11, pp. 2651–2661, Jun. 2005. ˆ19

[83] A. D. Leow, A. D. Klunder, C. R. Jack, A. W. Toga, A. M. Dale, M. A. Bernstein, P. J. Britson, J. L. Gunter, C. P. Ward, J. L. Whitwell, B. J. Borowski, A. S. Fleisher, N. C. Fox, D. Harvey, J. Kornak, N. Schuff, C. Studholme, G. E. Alexander, M. W. Weiner, and P. M. Thompson, "Longitudinal stability of MRI for mapping brain change using tensor-based morphometry." *Neuroimage*, vol. 31, no. 2, pp. 627–640, Jun. 2006, A.D.N.I. Preparatory Phase Study. ˆ19

[84] F. Fazekas, R. Kleinert, H. Offenbacher, R. Schmidt, G. Kleinert, F. Payer, H. Radner, and H. Lechner, "Pathologic correlates of incidental MRI white matter signal hyperintensities." *Neurology*, vol. 43, no. 9, pp. 1683–1689, Sep. 1993. ˆ20

[85] M. M. Breteler, J. C. van Swieten, M. L. Bots, D. E. Grobbee, J. J. Claus, J. H. van den Hout, F. van Harskamp, H. L. Tanghe, P. T. de Jong, and J. van Gijn, "Cerebral white matter lesions, vascular risk factors, and cognitive function in a population-based study: the Rotterdam Study." *Neurology*, vol. 44, no. 7, pp. 1246–1252, Jul. 1994. ˆ20

[86] C. D. Smith, D. A. Snowdon, H. Wang, and W. R. Markesbery, "White matter volumes and periventricular white matter hyperintensities in aging and dementia." *Neurology*, vol. 54, no. 4, pp. 838–842, Feb. 2000. ˆ20

[87] W. Wen, P. S. Sachdev, X. Chen, and K. Anstey, "Gray matter reduction is correlated with white matter hyperintensity volume: a voxel-based morphometric study in a large epidemiological sample." *Neuroimage*, vol. 29, no. 4, pp. 1031–1039, Feb. 2006. ˆ20

[88] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information." *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Apr. 1997. ˆ20, 24

[89] A. Roche, G. Malandain, X. Pennec, and N. Ayache, "The Correlation Ratio as a new similarity measure for multimodal image registration," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 1496, Jan. 1998, pp. 1115–. [Online]. Available: http://www.springerlink.com/content/pq3qc8v74k4pxue9/ ˆ20, 24

[90] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "Overlap invariant entropy measure of 3D medical image alignment," *Pattern Recognition*, vol. 32, no. 1, pp. 71–86, Jan. 1999. ˆ20, 24

[91] H. S. Choi, D. R. Haynor, and Y. Kim, "Partial volume tissue classification of multichannel magnetic resonance images - a mixel model," *IEEE Trans. Med. Imag.*, vol. 10, no. 3, pp. 395–407, Sep. 1991. ˆ20, 26

[92] J. Ashburner and K. Friston, "Multimodal image coregistration and partitioning–a unified framework." *Neuroimage*, vol. 6, no. 3, pp. 209–217, Oct. 1997. ˆ20, 27

[93] Y. Zhang, J. M. Brady, and S. M. Smith, "An HMRF-EM algorithm for partial volume segmentation of brain MRI - FMRIB Technical Report TR01YZ1," Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Tech. Rep., 2001. ˆ20, 27

[94] H. Hanyu, T. Asano, H. Sakurai, M. Takasaki, H. Shindo, and K. Abe, "Magnetization transfer measurements of the hippocampus in the early diagnosis of Alzheimer's disease." *J Neurol Sci*, vol. 188, no. 1-2, pp. 79–84, Jul. 2001. ˆ20

[95] A. C. G. M. van Es, W. M. van der Flier, F. Admiraal-Behloul, H. Olofsen, E. L. E. M. Bollen, H. A. M. Middelkoop, A. W. E. Weverling-Rijnsburger, R. G. J. Westendorp, and M. A. van Buchem, "Magnetization transfer imaging of gray and white matter in mild cognitive impairment and Alzheimer's disease." *Neurobiol Aging*, Nov. 2005. ˆ20

[96] W. M. van der Flier, D. M. J. van den Heuvel, A. W. E. Weverling-Rijnsburger, E. L. E. M. Bollen, R. G. J. Westendorp, M. A. van Buchem, and H. A. M. Middelkoop, "Magnetization transfer imaging in normal aging, mild cognitive impairment, and Alzheimer's disease." *Ann Neurol*, vol. 52, no. 1, pp. 62–67, Jul. 2002. ˆ20

[97] S. Hayasaka, A.-T. Du, A. Duarte, J. Kornak, G.-H. Jahng, M. W. Weiner, and N. Schuff, "A non-parametric approach for co-analysis of multi-modal brain imaging data: application to Alzheimer's disease." *Neuroimage*, vol. 30, no. 3, pp. 768–779, Apr. 2006. ˆ20

[98] S. E. Rose, F. Chen, J. B. Chalk, F. O. Zelaya, W. E. Strugnell, M. Benson, J. Semple, and D. M. Doddrell, "Loss of connectivity in Alzheimer's disease: an evaluation of white matter tract integrity with colour coded MR diffusion tensor imaging." *J Neurol Neurosurg Psychiatry*, vol. 69, no. 4, pp. 528–530, Oct. 2000. ˆ20

[99] A. Bozzao, R. Floris, M. E. Baviera, A. Apruzzese, and G. Simonetti, "Diffusion and perfusion MR imaging in cases of Alzheimer's disease: correlations with cortical atrophy and lesion load." *AJNR Am J Neuroradiol*, vol. 22, no. 6, pp. 1030–1036, 2001. ˆ20

[100] M. Bozzali, M. Franceschi, A. Falini, S. Pontesilli, M. Cercignani, G. Magnani, G. Scotti, G. Comi, and M. Filippi, "Quantification of tissue damage in AD using diffusion tensor and magnetization transfer MRI." *Neurology*, vol. 57, no. 6, pp. 1135–1137, Sep. 2001. ˆ20

[101] M. Rovaris, G. Iannucci, M. Cercignani, M. P. Sormani, N. De Stefano, S. Gerevini, G. Comi, and M. Filippi, "Age-related changes in conventional, magnetization transfer, and diffusion-tensor MR imaging findings: Study with whole-brain tissue histogram analysis." *Radiology*, vol. 227, no. 3, pp. 731–738, 2003. ˆ20

[102] N. Schuff, D. Amend, F. Ezekiel, S. K. Steinman, J. Tanabe, D. Norman, W. Jagust, J. H. Kramer, J. A. Mastrianni, G. Fein, and M. W. Weiner, "Changes of hippocampal N-acetyl aspartate and volume in Alzheimer's disease. A proton MR spectroscopic imaging and MRI study." *Neurology*, vol. 49, no. 6, pp. 1513–1521, Dec. 1997. ˆ20

[103] G. Tsai and J. T. Coyle, "N-acetylaspartate in neuropsychiatric disorders." *Prog Neurobiol*, vol. 46, no. 5, pp. 531–540, Aug. 1995. ˆ20

[104] M. J. Valenzuela and P. Sachdev, "Magnetic resonance spectroscopy in AD." *Neurology*, vol. 56, no. 5, pp. 592–598, Mar. 2001. ˆ20

[105] N. B. Walters, G. F. Egan, J. J. Kril, M. Kean, P. Waley, M. Jenkinson, and J. D. G. Watson, "In vivo identification of human cortical areas using high-resolution MRI: an approach to cerebral structure-function correlation." *Proc Natl Acad Sci U S A*, vol. 100, no. 5, pp. 2981–2986, Mar. 2003. ˆ20

[106] P. Batchelor, A. Castellano Smith, D. Hill, D. Hawkes, T. Cox, and A. Dean, "Measures of folding applied to the development of the human fetal brain." *IEEE Trans. Med. Imag.*, vol. 21, no. 8, pp. 953–965, Aug. 2002. ˆ20

[107] S. Ourselin, E. Bardinet, D. Dormont, G. Malandain, A. Roche, N. Ayache, D. Tandé, K. Parain, and J. Yelnik, "Fusion of histological sections and MR images: Towards the construction of an atlas of the human basal ganglia," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 2208, Jan. 2001, pp. 743–. [Online]. Available: http://www.springerlink.com/content/16ugttmw7k8mhuuy/ ˆ21

[108] A. Bardinet, S. Ourselin, D. Dormont, G. Malandain, D. Tandé, K. Parain, N. Ayache, and J. Yelnik, "Co-registration of histological, optical and MR data of the human brain," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 2488, Jan. 2002, pp. 548–555. ˆ21

[109] A. Pitiot, E. Bardinet, P. M. Thompson, and G. Malandain, "Piecewise affine registration of biological images for volume reconstruction." *Med Image Anal*, vol. 10, no. 3, pp. 465–483, Jun. 2006. ˆ21

[110] M. M. Chakravarty, G. Bertrand, C. P. Hodge, A. F. Sadikot, and D. L. Collins, "The creation of a brain atlas for image guided neurosurgery using serial histological data." *Neuroimage*, vol. 30, no. 2, pp. 359–376, Apr. 2006. ˆ21

[111] J. B. Maintz and M. A. Viergever, "A survey of medical image registration." *Med Image Anal*, vol. 2, no. 1, pp. 1–36, Mar. 1998. ˆ21

[112] J. V. Hajnal, D. L. G. Hill, and D. J. Hawkes, Eds., *Medical Image Registration*. CRC Press, 2001. ˆ21, 22

[113] B. Zitová and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, pp. 977–1000, 2003. ˆ21

[114] W. R. Crum, T. Hartkens, and D. L. G. Hill, "Non-rigid image registration: theory and practice." *Br J Radiol*, vol. 77 Spec No 2, pp. S140–S153, 2004. ˆ22

[115] R. P. Woods, S. T. Grafton, J. D. Watson, N. L. Sicotte, and J. C. Mazziotta, "Automated image registration: II. Intersubject validation of linear and nonlinear models." *J Comput Assist Tomogr*, vol. 22, no. 1, pp. 153–165, 1998. ˆ23

[116] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images." *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999. ˆ23

[117] J. Ashburner and K. J. Friston, "Nonlinear spatial normalization using basis functions." *Hum Brain Mapp*, vol. 7, no. 4, pp. 254–266, 1999. ˆ23

[118] R. Bajcsy and S. Kovačič, "Multiresolution elastic matching," *Comput. Vision Graph. Image Process.*, vol. 46, no. 1, pp. 1–21, 1989. ˆ23

[119] G. E. Christensen, R. D. Rabbitt, and M. I. Miller, "3D brain mapping using a deformable neuroanatomy." *Phys Med Biol*, vol. 39, no. 3, pp. 609–618, Mar. 1994. ˆ23

[120] P. A. Freeborough and N. C. Fox, "Modeling brain deformations in Alzheimer disease by fluid registration of serial 3D MR images." *J Comput Assist Tomogr*, vol. 22, no. 5, pp. 838–843, 1998. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9754126 ˆ23

[121] H. Lester, S. R. Arridge, K. M. Jansons, L. Lemieux, J. V. Hajnal, and A. Oatridge, "Non-linear registration with the variable viscosity fluid algorithm," in *Inf. Process. Med. Imag.*, ser. Lecture Notes in Computer Science, vol. 1613, Jan. 1999, pp. 238–. ˆ23

[122] W. R. Crum, C. Tanner, and D. J. Hawkes, "Anisotropic multi-scale fluid registration: evaluation in magnetic resonance breast imaging." *Phys Med Biol*, vol. 50, no. 21, pp. 5153–5174, Nov. 2005. ˆ23, 25

[123] J. P. Thirion, "Image matching as a diffusion process: an analogy with Maxwell's demons." *Med Image Anal*, vol. 2, no. 3, pp. 243–260, Sep. 1998. ˆ23

[124] J. Modersitzki, *Numerical Methods for Image Registration*. Oxford University Press, 2003. ˆ23

[125] C. J. Twining and S. Marsland, "Constructing an atlas for the diffeomorphism group of a compact manifold with boundary, with application to the analysis of image registrations," *Journal of Computational and Applied Mathematics*, 2008. ˆ23

[126] J. Ashburner, "A fast diffeomorphic image registration algorithm." *Neuroimage*, vol. 38, no. 1, pp. 95–113, Oct. 2007. ˆ23, 26

[127] M. Hernandez, M. N. Bossa, and S. Olmos, "Registration of anatomical images using geodesic paths of diffeomorphisms parameterized with stationary vector fields," in *Proc. IEEE 11th International Conference on Computer Vision*, M. N. Bossa, Ed., 2007, pp. 1–8. ˆ23

[128] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Non-parametric diffeomorphic image registration with the demons algorithm." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 10, no. Pt 2, 2007, pp. 319–326. [Online]. Available: http://www.springerlink.com/content/0uj2712ju7r554q1/ ˆ23

[129] M. Miller, A. Trouvé, and L. Younes, "Geodesic shooting for computational anatomy," *Journal of Mathematical Imaging and Vision*, vol. 24, no. 2, pp. 209–228, 2006. [Online]. Available: http://www.springerlink.com/content/9r82230441886375/ ˆ23

[130] S. Marsland and R. McLachlan, "A Hamiltonian particle method for diffeomorphic image registration." in *Inf. Process. Med. Imag.*, vol. 20, 2007, pp. 396–407. [Online]. Available: http://www.springerlink.com/content/x5q066610172r113/ ˆ23

[131] R. P. Woods, "Characterizing volume and surface deformations in an atlas framework: theory, applications, and implementation." *Neuroimage*, vol. 18, no. 3, pp. 769–788, Mar. 2003. ˆ24

[132] M. I. Miller, C. E. Priebe, A. Qiu, B. Fischl, A. Kolasny, T. Brown, Y. Park, J. T. Ratnanather, E. Busa, J. Jovicich, P. Yu, B. C. Dickerson, R. L. Buckner, and the Morphometry BIRN, "Collaborative computational anatomy: An MRI morphometry study of the human brain via diffeomorphic metric mapping." *Hum Brain Mapp*, Sep. 2008. ˆ24

[133] V. Arsigny, O. Commowick, X. Pennec, and N. Ayache, "A log-Euclidean framework for statistics on diffeomorphisms." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 9, no. Pt 1, 2006, pp. 924–931. [Online]. Available: http://www.springerlink.com/content/607206763v078397/ ˆ24

[134] N. Lepore, C. Brun, Y.-Y. Chou, A. Lee, M. Barysheva, X. Pennec, K. McMahon, M. Meredith, G. de Zubicaray, M. Wright, A. Toga, and P. Thompson, "Best individual template selection from deformation tensor minimization," in *Proc. 5th IEEE International Symposium on Biomedical Imaging*, 2008, pp. 460–463. ˆ24

[135] J. Ashburner, J. L. Andersson, and K. J. Friston, "Image registration using a symmetric prior–in three dimensions." *Hum Brain Mapp*, vol. 9, no. 4, pp. 212–225, Apr. 2000. [Online]. Available: http://www3.interscience.wiley.com/journal/71001030/abstract ˆ24

[136] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, pp. 567–585, 1989. ˆ24

[137] Y. Ge, J. M. Fitzpatrick, R. M. Kessler, M. Jeske-Janicka, and R. A. Margolin, "Intersubject brain image registration using both cortical and subcortical landmarks," in *Medical Imaging 1995: Image Processing*, M. H. Loew, Ed., vol. 2434. SPIE, 1995, pp. 81–95. [Online]. Available: http://link.aip.org/link/?PSI/2434/81/1 ˆ24

[138] S. Joshi and M. Miller, "Landmark matching via large deformation diffeomorphisms," *IEEE Trans. Image Process.*, vol. 9, no. 8, pp. 1357–1370, Aug. 2000. ˆ24

[139] P. Viola, "Alignment by maximization of Mutual Information," Ph.D. dissertation, Massachusetts Institute of Technology, 1995. ˆ24

[140] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey." *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 986–1004, Aug. 2003. ˆ24

[141] M. Holden, D. Hill, E. Denton, J. Jarosz, T. Cox, T. Rohlfing, J. Goodey, and D. Hawkes, "Voxel similarity measures for 3-D serial MR brain image registration," *IEEE Trans. Med. Imag.*, vol. 19, no. 2, pp. 94–102, Feb. 2000. ˆ24

[142] W. R. Crum, D. L. G. Hill, and D. J. Hawkes, "Information theoretic similarity measures in non-rigid registration." in *Inf. Process. Med. Imag.*, vol. 18, Jul. 2003, pp. 378–387. ˆ24

[143] A. Roche, G. Malandain, and N. Ayache, "Unifying maximum likelihood approaches in medical image registration," *International Journal of Imaging Systems and Technology*, vol. 11, no. 1, pp. 71–80, 2000. ˆ24

[144] B. Reddy and B. Chatterji, "An FFT-based technique for translation, rotation, and scale-invariant image registration," *IEEE Trans. Image Process.*, vol. 5, no. 8, pp. 1266–1271, 1996. ˆ24

[145] M. Mellor and M. Brady, "Non-rigid multimodal image registration using local phase," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 3216, Jan. 2004, pp. 789–796. [Online]. Available: http://www.springerlink.com/content/1ch4ecdje35yxetq/ ˆ24

[146] B. Jian, B. C. Vemuri, and J. L. Marroquin, "Robust nonrigid multimodal image registration using local frequency maps," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 3565, Jul. 2005, pp. 504–515. ˆ24

[147] J. P. Pluim, J. B. Maintz, and M. A. Viergever, "Image registration by maximization of combined mutual information and gradient information." *IEEE Trans. Med. Imag.*, vol. 19, no. 8, pp. 809–814, Aug. 2000. ˆ24

[148] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "F-information measures in medical image registration." *IEEE Trans. Med. Imag.*, vol. 23, no. 12, pp. 1508–1516, Dec. 2004. ˆ24

[149] C. Guetter, C. Xu, F. Sauer, and J. Hornegger, "Learning based non-rigid multimodal image registration using Kullback-Leibler divergence." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 8, no. Pt 2, 2005, pp. 255–262. ˆ24

[150] B. Fischer and J. Modersitzki, "Intensity-based image registration with a guaranteed one-to-one point match." *Methods Inf Med*, vol. 43, no. 4, pp. 327–330, 2004. ˆ25

[151] H. Johnson and G. Christensen, "Consistent landmark and intensity-based image registration," *IEEE Trans. Med. Imag.*, vol. 21, no. 5, pp. 450–461, 2002. ˆ25

[152] T. Hartkens, D. Hill, A. Castellano-Smith, D. Hawkes, C. Maurer, Jr., A. Martin, W. Hall, H. Liu, and C. Truwit, "Using points and surfaces to improve voxel-based non-rigid registration," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 2489, Jan. 2002, pp. 565–572. ˆ25

[153] J. Kybic and M. Unser, "Fast parametric elastic image registration," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1427–1442, Nov. 2003. ˆ25

[154] P. Cachier, J.-F. Mangin, X. Pennec, D. Rivière, D. Papadopoulos-Orfanos, J. Régis, and N. Ayache, "Multisubject non-rigid registration of brain MRI using intensity and geometric features," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, ser. LNCS, vol. 2208, Jan. 2001, pp. 734–. [Online]. Available: http://www.springerlink.com/content/jyjb1d0dwt8cy03y/ ˆ25

[155] E. Schreibmann and L. Xing, "Image registration with auto-mapped control volumes." *Med Phys*, vol. 33, no. 4, pp. 1165–1179, Apr. 2006. ˆ25

[156] D. Shen and C. Davatzikos, "HAMMER: hierarchical attribute matching mechanism for elastic registration." *IEEE Trans. Med. Imag.*, vol. 21, no. 11, pp. 1421–1439, Nov. 2002. ˆ25

[157] ——, "Measuring temporal morphological changes robustly in brain MR images via 4-dimensional template warping." *Neuroimage*, vol. 21, no. 4, pp. 1508–1517, Apr. 2004. ˆ25

[158] Z. Xue, D. Shen, and C. Davatzikos, "Determining correspondence in 3-D MR brain images using attribute vectors as morphological signatures of voxels." *IEEE Trans. Med. Imag.*, vol. 23, no. 10, pp. 1276–1291, Oct. 2004. ˆ25

[159] J. L. Lancaster, P. T. Fox, H. Downs, D. S. Nickerson, T. A. Hander, M. E. Mallah, P. V. Kochunov, and F. Zamarripa, "Global spatial normalization of human brain using convex hulls." *J Nucl Med*, vol. 40, no. 6, pp. 942–955, Jun. 1999. ˆ25

[160] P. M. Thompson and A. W. Toga, "A surface-based technique for warping three-dimensional images of the brain," *IEEE Trans. Med. Imag.*, vol. 15, no. 4, pp. 402–417, Aug. 1996. ˆ25

[161] D. L. Collins, G. L. Goualher, and A. C. Evans, "Non-linear cerebral registration with sulcal constraints," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, 1998. ˆ25

[162] M. Vaillant and C. Davatzikos, "Hierarchical matching of cortical features for deformable brain image registration," in *Inf. Process. Med. Imag.*, ser. Lecture Notes in Computer Science, vol. 1613, Jan. 1999, p. 182. ˆ25

[163] H. Lester and S. R. Arridge, "A survey of hierarchical non-linear medical image registration." *Pattern Recognition*, vol. 32, pp. 129–149, 1999. ˆ25

[164] J. A. Schnabel, D. Rueckert, M. Quist, J. M. Blackall, A. D. Castellano-Smith, T. Hartkens, G. P. Penney, W. A. Hall, H. Liu, C. L. Truwit, F. A. Gerritsen, D. L. G. Hill, and D. J. Hawkes, "A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 2208, Jan. 2001, pp. 573–. [Online]. Available: http://www.springerlink.com/content/0q3bfeunq4avrwdf/ ˆ25

[165] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images." *Neuroimage*, vol. 17, no. 2, pp. 825–841, Oct. 2002. ˆ25

[166] M. Droske and M. Rumpf, "A variational approach to nonrigid morphological image registration," *SIAM Journal on Applied Mathematics*, vol. 64, no. 2, pp. 668–687, 2004. ˆ25

[167] W. Lu, M.-L. Chen, G. H. Olivera, K. J. Ruchala, and T. R. Mackie, "Fast free-form deformable registration via calculus of variations." *Phys Med Biol*, vol. 49, no. 14, pp. 3067–3087, Jul. 2004. ˆ25

[168] R. Stefanescu, X. Pennec, and N. Ayache, "Grid powered nonlinear image registration with locally adaptive regularization." *Med Image Anal*, vol. 8, no. 3, pp. 325–342, Sep. 2004. ˆ25

[169] P. V. Kochunov, J. L. Lancaster, and P. T. Fox, "Accurate high-speed spatial normalization using an octree method." *Neuroimage*, vol. 10, no. 6, pp. 724–737, Dec. 1999. ˆ25

[170] H. Wang, L. Dong, J. O'Daniel, R. Mohan, A. S. Garden, K. K. Ang, D. A. Kuban, M. Bonnen, J. Y. Chang, and R. Cheung, "Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy." *Phys Med Biol*, vol. 50, no. 12, pp. 2887–2905, Jun. 2005. ˆ25

[171] A. Leow, C. L. Yu, S. J. Lee, S. C. Huang, H. Protas, R. Nicolson, K. M. Hayashi, A. W. Toga, and P. M. Thompson, "Brain structural mapping using a novel hybrid implicit/explicit framework based on the level-set method." *Neuroimage*, vol. 24, no. 3, pp. 910–927, Feb. 2005. ˆ25

[172] J. Sethian, *Level Set Methods and Fast Marching Methods.* Cambridge University Press, 1999. ˆ25, 27

[173] Z. Xue, D. Shen, and C. Davatzikos, "CLASSIC: consistent longitudinal alignment and segmentation for serial image computing." *Neuroimage*, vol. 30, no. 2, pp. 388–399, Apr. 2006. ˆ25, 27

[174] J. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, C. R. Maurer, R. M. Kessler, R. J. Maciunas, C. Barillot, D. Lemoine, A. Collignon, F. Maes, P. Suetens, D. Vandermeulen, P. A. van den Elsen, S. Napel, T. S. Sumanaweera, B. Harkness, P. F. Hemler, D. L. Hill, D. J. Hawkes, C. Studholme, J. B. Maintz, M. A. Viergever, G. Malandain, and R. P. Woods, "Comparison and evaluation of retrospective intermodality brain image registration techniques." *J Comput Assist Tomogr*, vol. 21, no. 4, pp. 554–566, 1997. ˆ25

[175] J. A. Schnabel, C. Tanner, A. D. Castellano-Smith, A. Degenhard, M. O. Leach, D. R. Hose, D. L. G. Hill, and D. J. Hawkes, "Validation of nonrigid image registration using finite-element methods: application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 22, no. 2, pp. 238–247, 2003. ˆ25

[176] W. R. Crum, L. D. Griffin, D. L. G. Hill, and D. J. Hawkes, "Zen and the art of medical image registration: correspondence, homology, and quality." *Neuroimage*, vol. 20, no. 3, pp. 1425–1437, Nov. 2003. ˆ25

[177] K. Amunts, A. Schleicher, U. Bürgel, H. Mohlberg, H. B. Uylings, and K. Zilles, "Broca's region revisited: cytoarchitecture and intersubject variability." *J Comp Neurol*, vol. 412, no. 2, pp. 319–341, Sep. 1999. [Online]. Available: http://www3.interscience.wiley.com/journal/63002079/abstract ˆ26

[178] P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. L. Goualher, D. L. Collins, A. Evans, G. Malandain, N. Ayache, G. E. Christensen, and H. J. Johnson, "Retrospective evaluation of intersubject brain registration." *IEEE Trans. Med. Imag.*, vol. 22, no. 9, pp. 1120–1130, Sep. 2003. ˆ26

[179] V. Noblet, C. Heinrich, F. Heitz, and J.-P. Armspach, "Retrospective evaluation of a topology preserving non-rigid registration method." *Med Image Anal*, vol. 10, no. 3, pp. 366–384, Jun. 2006. ˆ26

[180] J. He and G. E. Christensen, "Large deformation inverse consistent elastic image registration." in *Inf. Process. Med. Imag.*, vol. 18, Jul. 2003, pp. 438–449. ˆ26

[181] W. R. Crum, D. Rueckert, M. Jenkinson, D. Kennedy, and S. M. Smith, "A framework for detailed objective comparison of non-rigid registration algorithms in Neuroimaging," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 3216, Jan. 2004, pp. 679–686. [Online]. Available: http://www.springerlink.com/content/gh14cb1dupu0b3v8/ ˆ26

[182] W. R. Crum, O. Camara, D. Rueckert, K. K. Bhatia, M. Jenkinson, and D. L. G. Hill, "Generalised overlap measures for assessment of pairwise and groupwise image registration and segmentation." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 8, no. Pt 1, 2005, pp. 99–106. ˆ26, 27

[183] P. Kochunov, J. Lancaster, P. Thompson, A. W. Toga, P. Brewer, J. Hardies, and P. Fox, "An optimized individual target brain in the Talairach coordinate system." *Neuroimage*, vol. 17, no. 2, pp. 922–927, Oct. 2002. ˆ26

[184] P. Kochunov, J. L. Lancaster, P. Thompson, R. Woods, J. Mazziotta, J. Hardies, and P. Fox, "Regional spatial normalization: toward an optimal target." *J Comput Assist Tomogr*, vol. 25, no. 5, pp. 805–816, 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11584245 ˆ26

[185] K. Bhatia, J. Hajnal, B. Puri, A. Edwards, and D. Rueckert, "Consistent groupwise non-rigid registration for atlas construction," in *Biomedical Imaging: Macro to Nano, IEEE International Symposium on*, vol. 1, Apr. 2004, pp. 908–911. ˆ26

[186] S. Marsland and C. Twining, "Constructing diffeomorphic representations for the groupwise analysis of nonrigid registrations of medical images," *IEEE Trans. Med. Imag.*, vol. 23, no. 8, pp. 1006–1020, 2004. ˆ26

[187] S. Joshi, B. Davis, M. Jomier, and G. Gerig, "Unbiased diffeomorphic atlas construction for computational anatomy." *Neuroimage*, vol. 23 Suppl 1, pp. S151–S160, 2004. ˆ26

[188] J. Ashburner and K. J. Friston, "Computing average shaped tissue probability templates." *Neuroimage*, vol. 45, no. 2, pp. 333–341, Apr. 2009. ˆ26

[189] N. R. Pal and S. K. Pal, "A review on image segmentation techniques." *Pattern Recognition*, vol. 26, no. 9, p. 1277, 1993. ˆ26

[190] R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 6, pp. 641–647, 1994. ˆ26

[191] V. Grau, A. Mewes, M. Alcaniz, R. Kikinis, and S. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 447–458, Apr. 2004. ˆ26

[192] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, Jan. 1988. ˆ26

[193] A. Blake and M. Isard, *Active contours*. Springer-Verlag, 1998. ˆ26

[194] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active Shape Models - Their training and application," *Computer Vision and Image Understanding*, vol. 61, pp. 38–59, 1995. ˆ26

[195] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, 2001. ˆ26

[196] G. Calmon and N. Roberts, "Automatic measurement of changes in brain volume on consecutive 3D MR images by segmentation propagation," *Magnetic Resonance Imaging*, vol. 18, no. 4, pp. 439–453, May 2000. ˆ26, 27

[197] W. R. Crum, R. I. Scahill, and N. C. Fox, "Automated hippocampal segmentation by regional fluid registration of serial MRI: validation and application in Alzheimer's disease." *Neuroimage*, vol. 13, no. 5, pp. 847–855, May 2001. ˆ26

[198] J. C. Bezdek, L. O. Hall, and L. P. Clarke, "Review of MR image segmentation techniques using pattern recognition." *Med Phys*, vol. 20, no. 4, pp. 1033–1048, 1993. ˆ26

[199] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, 2002. ˆ26

[200] K. K. Leung, N. Saeed, K. Changani, S. P. Campbell, and D. L. G. Hill, "Spatio-temporal segmentation of rheumatoid arthritis lesions in serial MR images of joints," in *MMBIA*, 2006. ˆ26

[201] J. Ashburner and K. J. Friston, "Unified segmentation." *Neuroimage*, vol. 26, no. 3, pp. 839–851, Jul. 2005. ˆ27

[202] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm." *IEEE Trans. Med. Imag.*, vol. 20, no. 1, pp. 45–57, Jan. 2001. ˆ27

[203] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale, "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain." *Neuron*, vol. 33, no. 3, pp. 341–355, Jan. 2002. ˆ27

[204] K. V. Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based tissue classification of MR images of the brain," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 897–908, Oct. 1999. ˆ27

[205] J. Marroquin, B. Vemuri, S. Botello, E. Calderon, and A. Fernandez-Bouzas, "An accurate and efficient Bayesian method for automatic segmentation of brain MRI," *IEEE Trans. Med. Imag.*, vol. 21, no. 8, pp. 934–945, 2002. ˆ27

[206] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy, "Magnetic resonance image tissue classification using a partial volume model." *Neuroimage*, vol. 13, no. 5, pp. 856–876, May 2001. ˆ27

[207] W. M. Wells, III, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz, "Adaptive segmentation of MRI data," *IEEE Trans. Med. Imag.*, vol. 15, pp. 429–442, 1996. ˆ27

[208] O. Chapelle, P. Haffner, and V. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1055–1064, 1999. ˆ27

[209] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky, "Cortex segmentation: a fast variational geometric approach," *IEEE Trans. Med. Imag.*, vol. 21, no. 12, pp. 1544–1551, Dec. 2002. ˆ27

[210] S. M. Smith, "Fast robust automated brain extraction." *Hum Brain Mapp*, vol. 17, no. 3, pp. 143–155, Nov. 2002. ˆ27

[211] C. Fennema-Notestine, I. B. Ozyurt, C. P. Clark, S. Morris, A. Bischoff-Grethe, M. W. Bondi, T. L. Jernigan, B. Fischl, F. Segonne, D. W. Shattuck, R. M. Leahy, D. E. Rex, A. W. Toga, K. H. Zou, and G. G. Brown, "Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: effects of diagnosis, bias correction, and slice location." *Hum Brain Mapp*, vol. 27, no. 2, pp. 99–113, Feb. 2006. ˆ27

[212] K. Mardia, J. Kent, and J. Bibby, *Multivariate analysis*. Academic Press, 1979. ˆ28

[213] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, 8th ed. Iowa State University Press, 1989. ˆ28, 29

[214] R. Freund and W. Wilson, *Regression analysis: statistical modeling of a response variable.* Academic Press, 1998. ˆ29, 30

[215] W. G. Cochran and G. M. Cox, *Experimental designs.* Wiley New York, 1992. ˆ29

[216] D. J. Hand and M. J. Crowder, *Practical Longitudinal Data Analysis.* CRC Press, 1996. ˆ29

[217] P. Diggle, P. Heagerty, K.-Y. Liang, and S. L. Zeger, *Analysis of Longitudinal Data*, 2nd ed. Oxford University Press, 2002. ˆ29

[218] R. Henson and W. Penny, "ANOVAs and SPM," Wellcome Department of Imaging Neuroscience, University College London, UK, Tech. Rep., 2003. ˆ29

[219] J. N. Matthews, D. G. Altman, M. J. Campbell, and P. Royston, "Analysis of serial measurements in medical research." *BMJ*, vol. 300, no. 6719, pp. 230–235, Jan. 1990. ˆ29

[220] G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, Eds., *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods.* Chapman & Hall, 2008. ˆ29

[221] K. J. Friston, A. P. Holmes, and K. J. Worsley, "How many subjects constitute a study?" *Neuroimage*, vol. 10, no. 1, pp. 1–5, Jul. 1999. ˆ29

[222] J. A. Mumford and T. Nichols, "Simple group fmri modeling and inference." *Neuroimage*, vol. 47, no. 4, pp. 1469–1475, Oct. 2009. ˆ29

[223] K. J. Friston, K. E. Stephan, T. E. Lund, A. Morcom, and S. Kiebel, "Mixed-effects and fMRI studies." *Neuroimage*, vol. 24, no. 1, pp. 244–252, Jan. 2005. ˆ29, 33

[224] C. F. Beckmann, M. Jenkinson, and S. M. Smith, "General multilevel linear modeling for group analysis in fmri." *Neuroimage*, vol. 20, no. 2, pp. 1052–1063, Oct. 2003. ˆ29

[225] L. Frison and S. J. Pocock, "Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design." *Stat Med*, vol. 11, no. 13, pp. 1685–1704, Sep. 1992. ˆ29, 32

[226] A. J. Vickers, "The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study." *BMC Med Res Methodol*, vol. 1, p. 6, 2001. ˆ30

[227] ——, "Analysis of variance is easily misapplied in the analysis of randomized trials: a critique and discussion of alternative statistical approaches." *Psychosom Med*, vol. 67, no. 4, pp. 652–655, 2005. ˆ30, 32

[228] L. J. Frison and S. J. Pocock, "Linearly divergent treatment effects in clinical trials with repeated measures: efficient analysis using summary statistics." *Stat Med*, vol. 16, no. 24, pp. 2855–2872, Dec. 1997. ˆ30, 32

[229] C. E. McCulloch and S. R. Searle, *Generalized, Linear, and Mixed Models*. Wiley, 2001. ˆ33

[230] D. Glaser and K. Friston, *Variance Components*, 2nd ed. Academic Press, 2004, ch. 9. [Online]. Available: http://www.fil.ion.ucl.ac.uk/spm/doc/books/hbf2/pdfs/Ch9.pdf ˆ33

[231] H. Goldstein, *Multilevel Statistical Models*, 3rd ed. Kendall's Library of Statistics, 2003. ˆ33

[232] A. Skrondal and S. Rabe-Hesketh, *Generalized Latent Variable Modeling*. Chapman and Hall/CRC, 2004. ˆ33

[233] Y. Hochberg and A. Tamhane, *Multiple comparison procedures*. John Wiley & Sons, 1987. ˆ34

[234] T. Nichols and S. Hayasaka, "Controlling the familywise error rate in functional neuroimaging: a comparative review." *Stat Methods Med Res*, vol. 12, no. 5, pp. 419–446, Oct. 2003. ˆ34, 35, 36

[235] K. J. Rothman, "No adjustments are needed for multiple comparisons." *Epidemiology*, vol. 1, no. 1, pp. 43–46, Jan. 1990. ˆ35

[236] T. V. Perneger, "What's wrong with bonferroni adjustments." *BMJ*, vol. 316, no. 7139, pp. 1236–1238, Apr. 1998. ˆ35

[237] P. Westfall and R. Tobias, *Multiple Comparisons and Multiple Tests Using the SAS System*. SAS Publishing, 2000. ˆ35

[238] K. Friston, A. Holmes, K. Worsley, J. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: a general linear approach," *Human Brain Mapping*, vol. 2, pp. 189–210, 1995. ˆ35

[239] K. J. Worsley, A. C. Evans, S. Marrett, and P. Neelin, "A three-dimensional statistical analysis for cbf activation studies in human brain." *J Cereb Blood Flow Metab*, vol. 12, no. 6, pp. 900–918, Nov. 1992. ˆ35

[240] P. Westfall and S. Young, *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley-Interscience, 1993. ˆ36

[241] K. Worsley, "The geometry of random images," *Chance*, vol. 9, no. 1, pp. 27–40, 1996. [Online]. Available: http://www.math.mcgill.ca/keith/chance/chance.ps.gz ˆ36

[242] K. Worsley, S. Marrett, P. Neelin, A. Vandal, K. Friston, A. Evans *et al.*, "A unified statistical approach for determining significant signals in images of cerebral activation," *Human Brain Mapping*, vol. 4, no. 1, pp. 58–73, 1996. ˆ36

[243] J. Cao and K. Worsley, "The detection of local shape changes via the geometry of Hotelling's T^2 fields," *The Annals of Statistics*, vol. 27, no. 3, pp. 925–942, 1999. ˆ36

[244] R. Casanova, R. Srikanth, A. Baer, P. J. Laurienti, J. H. Burdette, S. Hayasaka, L. Flowers, F. Wood, and J. A. Maldjian, "Biological parametric mapping: A statistical toolbox for multimodality brain image analysis." *Neuroimage*, vol. 34, no. 1, pp. 137–143, Jan. 2007. ˆ36

[245] A. P. Holmes, "Statistical issues in functional brain mapping," Ph.D. dissertation, University of Glasgow, 1994. [Online]. Available: http://www.fil.ion.ucl.ac.uk/spm/doc/theses/andrew/ ˆ36

[246] S. Hayasaka, K. L. Phan, I. Liberzon, K. J. Worsley, and T. E. Nichols, "Nonstationary cluster-size inference with random field and permutation methods." *Neuroimage*, vol. 22, no. 2, pp. 676–687, Jun. 2004. ˆ36

[247] S. Hayasaka, A. M. Peiffer, C. E. Hugenschmidt, and P. J. Laurienti, "Power and sample size calculation for neuroimaging studies by non-central random field theory." *Neuroimage*, vol. 37, no. 3, pp. 721–730, Sep. 2007. ˆ36

[248] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B*, vol. 57, no. 1, p. 289, 1995. ˆ37

[249] J. D. Storey, "The positive false discovery rate: A bayesian interpretation and the q-value," *The Annals of Statistics*, vol. 31, no. 6, pp. 2013–2035, Dec. 2003. [Online]. Available: http://www.jstor.org/stable/3448445 ˆ37

[250] C. R. Genovese, N. A. Lazar, and T. Nichols, "Thresholding of statistical maps in functional neuroimaging using the false discovery rate." *Neuroimage*, vol. 15, no. 4, pp. 870–878, Apr. 2002. ˆ37

[251] J. R. Chumbley and K. J. Friston, "False discovery rate revisited: Fdr and topological inference using gaussian random fields." *Neuroimage*, vol. 44, no. 1, pp. 62–70, Jan. 2009. ˆ37

Chapter 2

Permutation Testing

We review the theory of permutation testing for general linear models, providing new insights into the relationships between several different strategies which have been proposed. Extensions are discussed, including non-standard statistics and multivariate data, which are employed in chapter 4. An important theoretical contribution is a thorough exploration of a particular strategy which aims to achieve an exact test for a general linear model by performing the permutation on an exchangeable lower dimensional set of transformed residuals, before transforming back to the original data space. In particular, we connect this transformed-residual permutation test more clearly to the traditional strategies, and propose new alternative transformations for its residuals. The main practical novelty in this chapter is a set of Monte Carlo simulations comparing the different permutation test strategies in a range of challenging situations. The two new transformed-residual approaches and one variant on a traditional method have not previously been evaluated in the literature. Further experiments focus on important related aspects, including an investigation into different classes of permutations from which to sample.

2.1 Introduction

Parametric statistical tests employ a certain parametrised mathematical model for the distribution of a test statistic under the null hypothesis, usually relying on an assumption of normally distributed errors. Non-parametric techniques have been developed which use resampling or randomisation methods [1] to empirically estimate test statistic null distributions, removing the need for parametric assumptions. We distinguish two main alternatives: bootstrap methods [2] use resampling with replacement from the data (or more generally, from some estimate of the unobservable errors), to approximate the distribution of a parameter or statistic; randomisation or permutation methods [3] use re-ordering of the data (or estimated errors) to derive the permutation null distribution of a statistic. Good [3, 73] recommends permutation testing if the parametric distribution is unavailable or uncertain, and if exchangeability can be assumed; he suggests bootstrap methods are only necessary if neither parametric nor permutation methods are suitable, or if a confidence interval is required on a statistic which is not a parameter of the distribution. Permutation testing will be the focus of this chapter.

By relaxing the requirements on the test statistic, these testing methods have the potential to employ statistics which are more powerful, more robust, or more widely applicable. Randomisation tests provide exact control of type I error in some situations, and approximate control in general. Importantly for neuroimaging, randomisation tests can be easily extended to control family-wise error in multiple testing scenarios [4, 5]. The main drawback of resampling or randomisation tests are their significant computational cost, which has become less important with increasing computing power.

To date, much of the theory and implementation of randomisation testing has focussed on simple situations where the test is exact. For more general models, approximate permutation tests have been proposed which should still control false-positives.

2.2 Basic concepts

In the standard general linear model, $y = Xb + \varepsilon$, the elements of the error vector are assumed to be independently and identically distributed (IID) as a zero-mean Gaussian with a certain variance. This assumption endows the test statistic (derived from the likelihood ratio test, as described in appendix A.4.4) with a known parametric (F) distribution under the null hypothesis. The use of this distribution to produce p-values, confidence intervals, etc. is based on the classical statistical concept that if the null hypothesis were true, the distribution would accurately describe the behaviour over a large number of theoretical repetitions of the experiment (or observation). The fundamental idea behind the class of non-parametric statistical method considered here, is to replace this concept of theoretical repetitions with an empirical evaluation of the statistic's null distribution over a number of practical repetitions. These repetitions should be generated in a way which would be consistent with the assumptions, including the null hypothesis.

An example should help to clarify the concept. Consider the problem of testing whether two groups have equal mean. The standard two-sample t-test uses the distribution of the difference in sample means divided by the estimated standard error of this difference, which is t-distributed under the IID Gaussian assumption. For the randomisation testing equivalent, under the assumption that the two groups are random samples from two populations with the same mean we are justified to randomly reallocate the group labels under the null hypothesis [1]. If we compute the difference in means for all $(n_1+n_2)!/n_1!n_2!$ allocations of the data to two groups (with the original sizes, n_1 and n_2) we have an estimate of the null distribution which can be used to test the significance of the originally observed difference in means. If the real labelling's difference in means lies below the 5th percentile or above the 95th percentile of the randomisation distribution then we reject the null hypothesis at the 10% level. More precisely, we can assign a p-value equal to the proportion of the randomisation distribution as-or-more extreme than the original observation (with the original counted as part of the randomisation distribution). Critical values of the statistic for a given α can be similarly determined, and both one- or two-tailed alternative hypotheses are easily handled.

The randomisation test is exact, in the sense that its type I error under repeated

experiments where the null hypothesis is true will have an expected value exactly equal to the chosen α. For moderately large group sizes the exhaustive set of randomisations will be too large to evaluate in practice, but a random sampling from this set is likely to produce a very similar result. In fact, the randomisation test using the random subset can still be considered to be exact in some sense, see p.15 of Manly et al. [1].

In cases where the data are not randomly sampled, or randomly allocated in an experimental design, the test can still be justified under 'weak distributional assumptions' [4], for example that the distributions have the same shape. Note that this is a stronger assumption than simple equality of population means (in particular, the above test is sensitive to differences in group variances as well as differences in mean), but still a considerably weaker assumption than IID normality. Some authors distinguish between 'randomisation tests' and 'permutation tests' where the former have a justification in terms of random sampling or experimental design, while the latter are justified in terms of an assumed exchangeability [4]. We have attempted to follow this distinction above, but in the remainder of this work, it will have little relevance, and we will refer exclusively to permutation testing.

A second example leads towards the more general situation. Consider a simple regression problem, where the null hypothesis is that there is no (linear) relationship between the dependent variable y and the regressor x. If the data can be assumed to be exchangeable, then computing correlation coefficients for the original and permuted data will again allow us to derive a p-value for the observed correlation coefficient from its relative position in the permutation distribution.

Again, the assumption of exchangeability can be made either because the data are randomly sampled (from a bivariate population of independent x and y), or from a designed experiment (with the y values acquired after the x values were randomly assigned). Manly states that with observational studies, the justification is weaker and requires the null hypothesis to be that the x and y values are 'unrelated' [1], which might suggest that the errors should be IID (not necessarily Gaussian), in the sense that dependency or heteroscedasticity within the errors would mean that any relationship between the errors and the explanatory variable would invalidate the exchangeability of the data. In fact, the IID property is sufficient for exchangeability, but not necessary — for example, compound symmetric errors[1] are dependent but exchangeable [6]. For more on the concept of exchangeability see [6, 7]. Among the most useful results from Commenges [7] are: (i) a matrix M is exchangeable if and only if $SMS^T = M$ for all permutation matrices S; (ii) an orthonormal basis for the space of exchangeable matrices is given by[2] $\bar{1}_n = 1_{n\times n}/n$ and $I - \bar{1}_n$, (iii) data with a constant mean and exchangeable covariance matrix has 'second-moment exchangeability', which implies complete exchangeability for normally distributed data, and may lead to useful approximate exchangeability for more general distributions.

At this point, the above examples can be used to highlight two important aspects of permutation testing, by noting that a two-sample t-test is equivalent to a simple regression

[1] A compound symmetric covariance matrix has a constant diagonal, and a possibly different constant value everywhere else.

[2] These matrices are respectively the projection matrix P and residual forming matrix R for a design consisting only of a constant term: $X = 1_{n\times 1}$.

with a dummy variable indicating group membership.

2.2.1 Data or design permutation

In the two-group example, we talked of relabelling the groups, which can be thought of as permuting the group indicator variable; while in the correlation example, we talked of permuting the data. In fact, for any test statistic that depends only on the pairing of dependent and independent variables, and not on the order of the pairs as such, these interpretations are equivalent. Consider a permutation matrix S,[3] whose action is to shuffle the rows of a matrix which it premultiplies. All such S can be derived from permuting the rows (or columns) of an identity matrix, and are orthogonal $S^TS = I = SS^T$. Now, in the general linear model, with residual-forming matrix $R = I - X(X^TX)^+X^T$; after permuting the data, the sum of squares (from which, for full and reduced models, the t- or F-statistic is derived) is given by $(Sy)^TR(Sy) = y^TS^TRSy = y^TR_Sy$ with:

$$\begin{aligned} R_S &= S^TS - S^TX(X^TX)^+X^TS \\ &= I - S^TX(X^TSS^TX)^+X^TS \\ &= I - S^TX\left((S^TX)^T(S^TX)\right)^+(S^TX)^T \end{aligned}$$

where the final expression can immediately be seen to be the residual-forming matrix from a permuted design S^TX — i.e. permuting the data by S or the design by S^T are equivalent. This can also be seen in terms of the estimated parameters, where

$$\begin{aligned} \hat{B}^S &= X^+SY \\ &= (X^TX)^+X^TSY \\ &= (X^TSS^TX)^+X^TSY \\ &= ((S^TX)^T(S^TX))^+(S^TX)^TY \\ &= (S^TX)^+Y \end{aligned}$$

Trivially, we can also see that permuting both design and data by the same permutation matrix S has no effect on the model.

2.2.2 Choice of statistic

The second point to draw from the examples is that in the two-group example we suggested to use the difference in mean as the statistic, while in the regression, we chose the correlation coefficient. A related question which naturally arises when comparing a permutation test to a parametric version is whether the permutation test should use the same statistic as the parametric test, but without using the (assumed) parametric distribution. I.e. in this example, is there any advantage to using the t-statistic in place of the difference of means or correlation coefficient? This question is motivated by the fact that common parametric statistics often have some form of optimality, e.g. as described in ap-

[3] We avoid the letter P due to its association with projection matrices in linear modelling.

pendix A.4, t- and F-statistics derive from the generalised likelihood ratio test and so are optimal (under the parametric assumptions!) according to the Neyman-Pearson lemma. In these particular examples, it can be shown that all three statistics are in fact equivalent. Firstly, the difference in means is simply b_1 in a regression model with a constant (corresponding to b_0) and a binary variable coding group. Secondly, we note that any monotonic transformation of the statistic used for the permutation test will preserve the relative ordering of the transformed statistics in the permutation distribution (including the original) so the p-values (or critical values) which derive from the original statistic's relative position will remain unchanged. The t-statistic and (partial) correlation coefficient ρ satisfy [8]:[4]

$$t = \rho\sqrt{\frac{DF_E}{1-\rho^2}}$$
$$\rho = t\frac{1}{\sqrt{DF_E + t^2}}$$

which are monotonic functions as required. It remains to show that the t-statistic and b_1 are permutationally equivalent. They clearly have the same sign, so it would suffice to consider the relationship between $F = t^2$ and b_1^2. It is instructive to derive this as a special case of the relationship between a contrast of interest $c^T b$ and the corresponding F-statistic for a general linear model. Firstly, we note that the numerator and denominator degrees of freedom are not affected by permutation of the data, so F is monotonically related to SS_H/SS_E. Since $SS_R = SS_E + SS_H$, $SS_R/SS_H = 1 + SS_E/SS_H$ and so SS_H/SS_R is also monotonically related to F, as is SS_R/SS_E, so we can consider the ratio of any pair of the three sums of squares. From appendix A.4.4 and equation (A.19) we have

$$SS_H = b^T C(C^T(X^T X)^+ C)^+ C^T b,$$
$$SS_E = y^T R y,$$
$$SS_R = y^T R_0 y,$$

and we can also write $SS_E = y^T(I-P)y = y^T y - y^T P y = y^T y - b^T X^T X b$.

Now, we first note that $X^T X$ is invariant under permutation of X, and that in the case of a t-contrast $C^T(X^T X)^+ C$ and $C^T b$ will be scalar, so SS_H is indeed permutationally equivalent to $(C^T b)^2$. Similarly, $y^T y$ is invariant under permutation, and it may initially appear that $b^T X^T X b$ will be permutationally equivalent to $(C^T b)^2$, however this quadratic form in the vector b depends on the nuisance as well as interest elements of b, and therefore the relationship is potentially non-monotonic. Similarly, we argue that there is no reason in the general case to expect $SS_R = y^T R_0 y$ to have a monotonic relationship with $(C^T b)^2$. However, simple regressions with or without a constant are interesting special cases. With no constant term, the reduced model is the null matrix, and the residual-forming matrix is an identity, giving obvious permutational independence to SSR. With a constant,

[4]For completeness, note that the multiple correlation or squared coefficient of determination and the F-statistic are also similarly related: $F = (R^2/(1-R^2))DF_E/DF_H$ and $R^2 = F/(F + DF_E/DF_H)$.

$R_0 = I - 1_{n\times 1}1_{n\times 1}^+$, and $y^T S^T R_0 S y = y^T R_0 y$ is also permutationally invariant, thanks to the fact that permutation has no effect on the vectors of ones $1_{n\times 1}$.

Having established that $C^T b$ is not permutationally equivalent to its corresponding t- or F-statistic in the general case, the question arises as to whether either is preferable. Kennedy and Cade [9] argue that a permutation test should use a pivotal statistic, such as t, F or a correlation coefficient, and not an element of b. They demonstrate empirically unacceptable type I error when using b under a particular permutation testing strategy ('Shuffle-X', discussed in section 2.4). They also refer to related work on the importance of pivotal statistics in Monte Carlo and bootstrap methods.

A pivotal statistic is one whose sampling distribution is independent of unknown parameters [10].[5] For example, in testing a mean from a population with unknown variance, Student's t-statistic follows a t-distribution independent of the population mean or variance. This property allows the distribution to be computed or tabulated for given sample sizes (degrees of freedom). Clearly the ease of parametric representation or tabulation is of no interest in nonparametric permutation testing, however, the independence from unknown parameters can be a useful property for a permutation test statistic.

Nichols and Holmes [5] suggested that 'virtually any statistic' was suitable for permutation testing, but they preferred 'more pivotal' statistics to un-normalised ones such as $C^T b$. An additional motivation for pivotal statistics arises when using the permutation distribution of the image-wise maximum to correct for multiple comparisons (section 2.3).

2.2.3 Transformations of the data

Having emphasised the invariance of permutation inference to monotonic transformations of the test statistic, it seems important to clarify that permutation tests are not invariant to monotonic but nonlinear transformations of the actual data. For example, log-transforming strictly positive values (as will be relevant in chapter 4) can change the relative ordering of the statistics (using the same set of permutations) and hence the p-values obtained. However, Commenges [7] notes that identical component-wise nonlinear transformations (such as the log-transform) do preserve exchangeability, so will be valid for general permutation tests. A related point is the importance of the choice of function in multivariate combining function approaches, which are discussed briefly in section 2.3.4.

This is in contrast to non-parametric tests based on ranks [11], where the invariance to monotonic transformations does extend to the data. Manly points out on p.15 [1] that some standard non-parametric tests can be seen as randomisation tests where the dependence only on the relative ordering allows the complete randomisation distribution to be enumerated for particular designs and sample sizes. It is widely known that rank-based nonparametric methods tend to be less powerful than their parametric counterparts (given that the assumptions of the latter hold), Edgington states that permutation tests can avoid the loss of power from the rank-transformation [12], which is indirectly[6] supported

[5] A more mathematically complete definition can be found in [8].

[6] There seem not to be any direct comparisons of permutation testing with and without rank-transformation in the standard texts.

by Monte-Carlo studies comparing parametric and permutation tests [1, 5].

2.3 Family-wise error control with permutation testing

In section 1.6.4, we discussed the multiple testing problem inherent in voxel-wise statistical analysis of imaging data, and explained how results from Random Field Theory could be used to control the chance of false positives occurring in any of the analysed voxels (the family-wise error rate or FWE). This method is based on approximating the null distribution of the maximum of the voxels' statistics. Analogous to the way standard permutation testing replaces parametric assumptions about the null hypothesis with non-parametric estimation of an empirical null distribution, in the case of multiple tests we can naturally estimate the maximum distribution by simply recording the maximum statistic over the voxels for each permutation. We can then directly use percentiles from this permutational maximum distribution as critical thresholds, or use the ranking of the observed voxel-wise statistics within it to assign FWE corrected p-values.

Because permutation-based FWE control removes the need for (often approximate) closed-form expressions, it has the major advantage of allowing a much broader range of statistics. The types of statistic that should be valid for permutation methods in the multiple-testing situation are unchanged from the standard case. However, an additional consideration arises with multiple tests: it may be possible for the FWE to be accurately controlled overall while different voxels may have different sensitivities and specificities [5]. For example, consider an image which contains a region of highly variable voxels,[7] these voxels will be more likely to produce high values under the permuted labellings, and the maximum distribution will tend to be based more heavily on these voxels; less variable voxels may then be less likely to be found significant in comparison with this distribution. Voxels with true null hypotheses will have lower specificity with increasing variance, while voxels with true alternative hypotheses will have lower sensitivity with decreasing variance. For this reason, it may be desirable to use statistics that approximately have a common voxel-wise null distribution, such as the standard pivotal parametric statistics [5].

Using pivotal statistics such as t instead of un-normalised statistics like GLM contrasts $c^T b$, will result in more homogeneous null distributions across voxels. However, some heterogeneity may remain, for example variable skew can persist after variance has been standardised [13]. For this reason, Pantazis et al. [13] suggest replacing the maximum statistic approach with a minimum p-value one: at each voxel, the permutation distribution of that voxel can be used to derive an uncorrected non-parametric p-value, comparison of each voxel's p-value with the permutation distribution of the minimum of these p-values over all voxels can be used to control FWE. Since each uncorrected p-value should be uniformly distributed under the null hypothesis (without requiring Gaussian or other parametric assumptions), all voxels have a common null distribution and hence specificity is uniform across the image. However, there are two disadvantages with this ap-

[7] The example is a realistic one for voxel-based morphometry, where the accuracy of inter-subject registration will tend to be lower at structures which are more challenging to register and/or less biologically consistent across subjects.

proach. Firstly, in practical terms, it has very high memory requirements, since the entire permutation distribution consisting of $N_v \times N_p$ statistics is required, where there would typically be of the order of $N_v = 10^6$ voxels and $N_p = 10^3$ or 10^4 permutations.[8] Secondly, and perhaps more importantly, is the disadvantage that the non-parametric uncorrected p-values have a discreteness which is detrimental to the sensitivity of the FWE corrected results. More precisely, the permutation-based p-values are all multiples of $1/N_p$. If a large number L of the permutations achieve the lowest possible uncorrected p-value, then the corrected p-values can be no more significant than L/N_p, hence potentially requiring very large numbers of permutations [13].

2.3.1 Step-down FWE control

Step-wise methods have been proposed that offer more powerful control of FWE [14]. For example, a step-down modification of the standard Bonferroni procedure takes into account the fact that once we have declared the largest statistic to be significant (at the corrected level), the problem of controlling FWE over the remaining tests has multiplicity reduced by one, and so on. This is Holm's method, which compares the increasingly ordered p-values $p_{(i)}$ to $\alpha_0/(N - i + 1)$, stopping when they become larger, and declaring all smaller p-values significant. Nichols and Hayasaka found that such step-wise procedures had little advantage over the standard Bonferroni method for neuroimaging data, and that maximum-based random field theory or permutation testing approaches were significantly better [5].

However, in the context of maximum-based permutation methods for controlling FWE, it is possible that a similar step-down procedure could partially address the problem of non-uniform sensitivity due to heterogeneous null distributions [4]. Since we seek the permutation distribution of the maximum statistic under the null hypothesis, we can argue that voxels for which we reject the null hypothesis under the current estimate of this distribution should be excluded from a subsequent re-estimation of the maximum distribution. This should therefore mean that some of the highly variable (or skewed, etc.) voxels will be removed from the maximum distribution which is eventually used to test the less variable voxels, increasing the sensitivity of the latter. We cannot hope to achieve uniform sensitivity and specificity to the extent of the minimum p-value approach, in part because many of the highly variable voxels might not have high values in the original labelling and hence will not be excluded in the step-down procedure. However, the step-down approach does not suffer from the discreteness problem of the minimum p-value approach. An application with potential to benefit from this is mentioned in section 2.6.1.

Theoretically, step-down procedures cannot be less powerful than single-step methods, so they would seem a desirable option. The main draw-back is computational complexity. The simplest way of implementing an FWE-controlling step-down permutation test is to

[8] It is possible to determine the uncorrected p-values for the original labelling in an efficient manor by simply counting the number of times the original statistic is exceeded by the permuted versions; however, the minimum p-value distribution requires the equivalent of uncorrected p-values for all permutations, not just the original.

perform repeated runs of a standard permutation test, removing all of the significant voxels at each run [15]. This kind of blocked step-down test is not particularly time-efficient, and, will be slightly less powerful than a method which removes voxels individually. To remove individual voxels at a time, however, the complete permutation distribution must be stored. An interesting alternative has been proposed by Belmonte and Yurgelun-Todd [16] that records not only the values of the maxima in each permutation, but also their locations, and the values and locations of the next largest statistic, and so on for a certain number of N_r 'reserves' largest values, where N_r can be small compared to the number of permutations. This approach allows voxels to be removed from this partial maximum distribution and efficiently replaced with reserves. It is possible for a particular labelling to have all its stored reserves exhausted. This problem is particularly likely to occur with very smooth data, since voxels which are high in the original labelling will tend to be high or low together in other permutations due to the preservation of the spatial correlation that arises from permuting all voxels in the same way. Hence the set of reserves in a particular labelling in the partial permutation distribution might all come from the same cluster, and once this cluster has been found significant, the step-down procedure will have removed all the information about this labelling stored in the partial distribution. The solution Belmonte implemented in the AFNI software that accompanied the article (not described in the published paper itself) is to drop these exhausted labellings from the distribution, and to continue to evaluate the other voxels against a partial permutation distribution which has a reduced number of permutations as well as a reduced number of voxels. As long as the number of permutations is initially large, and the number of exhausted permutations remains fairly small, the approach could still be more powerful than a single-step equivalent, though it is important to note that it may also be less powerful, unlike a true step-down procedure.

We close this section by discussing a single-step procedure which is related in the sense that it also attempts to derive the maximum distribution from only null data. In a functional MEG study, Chau et al. [17] proposed that the permutation distribution of the maximum be derived only from some separate rest periods, by arbitrarily labelling (fake) activation periods within them, and then flipping the rest/activation signs in a group-wise one-sample t-test. The derived maximum distribution is then used with the correctly labelled rest/activation data, the hope being that this resting-state maximum distribution will not be broadened by the presence of voxels with true alternative hypotheses, hence delivering higher power. However, it is possible that differences such as larger variance or skew present in the genuine distribution compared to the rest/fake distribution could increase the false positive rate. In [17], this risk was minimised by using the same amount of data for the null distribution computation as was present in the main experiment. This need for symmetry between the experimental and null data makes the technique difficult to apply in more general contexts relevant to structural imaging. For example, it would not be valid to take only the data which has constant values of a categorical interest covariate and make a fake interest covariate within this set (E.g. in a two-sample comparison of controls vs. patients, splitting the controls into two new halves) to derive a maximum

distribution for use with the original design.

2.3.2 Non-standard statistics

Permutation-testing frees us from the need to know analytically the null-distribution of the statistic or of the maximum of a field of such statistics, since we can approximate this by the permutation distribution. It is therefore possible to consider other statistics, with fairly relaxed assumptions about their properties. In particular, we note that multivariate statistics are very easily handled. Section A.4.4, presents Rao's F-approximation for Wilks' Λ statistic:

$$F = \frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} / \frac{\nu_1}{\nu_2},$$
$$\Lambda = \frac{|Y^T R Y|}{|Y^T R_0 Y|}$$

The degrees of freedom and s depend only on terms such as the dimensionality of the data and hypothesis which are constant under permutation. The powers involved in both the F-approximation and the definition of Λ are monotonic, so we can equivalently base a permutation test directly on Λ^{-1}, i.e. the ratio of the determinants of the reduced- and full-model sums of squares and products matrices.[9] Hotelling's T^2 statistic, applicable to one- or two-sample tests of multivariate means, is also derived from the likelihood ratio [18], and is monotonically related to Wilks' Λ by

$$\Lambda^{2/n} = \left(1 + \frac{T^2}{n-1}\right)^{-1}, \tag{2.1}$$

and to the squared Mahalanobis distance by $T^2 \propto (\bar{x} - \mu)^T S^{-1} (\bar{x} - \mu)$, for sample mean $\bar{x}$, hypothesised population mean μ and sample covariance matrix S. Hence, all three of these common multivariate statistics are permutationally equivalent.

Robust statistics, i.e. statistics which aim to be less sensitive to outliers or to violations of the standard assumptions, usually lack simple parametric distributions. Subject to computational demands, they can be easily handled in the permutation testing framework. For example, Cade and Richards [19] explore permutation tests for least absolute deviation regression, where parametric testing can be problematic. Brammer et al. [20] proposed testing the median in place of the mean for greater robustness in permutation-based analysis of functional imaging data. Rorden et al. [21] investigated further generalised rank order statistics, with potential benefits for skewed distributions or other data for which the mean (or median) might not be the best measure of central tendency.

An interesting and successful example of an unusual statistic for which the parametric distribution is unavailable is the smoothed-variance pseudo t-statistic [22]. At low degrees of freedom, the estimated variance is noisier and/or rougher than the (usually anatomically contiguous) signal of interest. Since the variance is only needed in a permutation test to

[9] The reciprocal is taken simply to give a statistic for which large values indicate departure from the null hypothesis, for easier interpretation.

normalise the signal so as to achieve a more pivotal statistic (as discussed above), the denominator of the t-statistic can be spatially smoothed (in addition to any smoothing of the original data). This has been found to increase sensitivity compared to permutation testing of standard t-statistics [4].

2.3.3 Extent-based and related statistics

Following on from the previous section, we observe that the maximum distribution method of FWE control is very well suited to statistics that aim to capture information other than simple voxel-wise significance. Stemming from the intuition that a cluster of adjacent voxels with large effects is more likely to be biologically significant than more random-appearing isolated voxels, approaches have been developed to determine statistical significance based on the size of clusters that exceed a pre-specified threshold [23, 24]. Such cluster-size tests are more sensitive to weaker but spatially broader effects [25]. It is a simple matter to implement a cluster-size test in a permutation framework [22]; it is only necessary to record the maximum cluster size instead of the maximum statistic value. It is also possible to derive random field theory results for cluster size inference [24]. Imaging data may exhibit non-stationary smoothness, i.e. some regions of the image have spatially smoother residuals than others. Large clusters are more likely in smoother regions, which hence could admit false positives if a single average smoothness were assumed to apply everywhere in the image. Both random field theory and permutation-based cluster size inference can be extended to handle this case [26].

Either cluster extent or peak height could be indicative of an effect, and clusters which are both large and intense would provide the strongest indication, it is therefore desirable to base inference on some combination of these aspects, rather than considering either one alone. It is here that permutation testing really comes to the fore. A parametric test for combined cluster height and size has been proposed based on their bivariate joint distribution [27], but this necessitates an approximation and some strong assumptions [28]. Bullmore et al. [29] proposed that the two aspects be combined by summing the statistic values over the supra-threshold clusters, giving a measure known as cluster-mass, which does not have a known parametric distribution, but has been found to perform very favourably compared to standard alternatives [28, 29]. The above approaches are all based on the concept of a supra-threshold cluster, and hence require a cluster-defining threshold to be specified. The arbitrary choice of such threshold has no impact on the validity of the inference, though this arbitrariness is often disliked by practitioners. More importantly, while two different thresholds would both lead to valid results, their findings could differ in scientifically significant ways, causing problems for the interpretation. The permutation testing framework can again help here, by generalising the cluster-mass approach. Instead of setting a threshold and summing supra-threshold values, 'threshold-free cluster enhancement' [30] assigns a value to each voxel based on the sum of supporting sections beneath it, where the sections are defined over a range of heights from zero to the value at that voxel. As well as removing the need for an arbitrary threshold, the method has the advantage that distinct maxima are preserved, and can be interpreted within any

significant 'clusters' found through permutation testing.

2.3.4 Multivariate combining functions

We now briefly present an approach which has extremely wide applicability — deserving of more discussion than is given here — the use of combining functions for multivariate permutation testing [31]. In essence, separate 'partial' tests for the m multivariate components are combined into a single summary statistic, for example, p-values for partial tests $\{p_i\}_{i=1}^m$ may be amalgamated using Fisher's combining function $-2\sum_i \log p_i$. The permutation distribution of these new statistics can then be used to derive overall p-values. A combining function approach based on p-values provides useful scale-invariance (since each partial test's p-values are uniformly distributed on $[0, 1]$ under the null hypothesis) and implicitly accounts for correlation between the partial tests [31].

The multivariate combining function approach can be used to jointly test cluster extent and height [28]. Indeed cluster-mass [29] can be seen as a form of combining function [28].[10] Different combining functions can be used to obtain different properties, for example Tippet's combining function $1 - \min_i \log p_i$ is sensitive to situations in which one partial test is significant, but gains no additional sensitivity from greater significance of other non-extremal partial tests. For combining cluster size and peak height, Tippet's combining function is analogous to the parametric test discussed above [27], while cluster-mass is sensitive to simultaneously large and intense clusters but not necessarily significant for clusters which would be significant for either size or height alone, and Fisher's combining function is a compromise between these two extremes [28]. Some other combining functions are discussed in [32] and [31]. Hayasaka and Nichols proposed that the merits of different combining functions could be jointly capitalised on through the use of a second-level 'meta-combining' function of the combining functions [28].

In the context of controlling FWE, the combining function approach can be seen as a generalisation of the minimum p-value permutation test of Pantazis et al. [13] (described in section 2.3). Where Pantazis obtained permutation-based p-values at each voxel, and considered the permutation distribution of the minimum over the image, Pesarin's approach obtains multiple partial p-values at each voxel, combines these, and then considers the permutation distribution of the image-minima. This method has been employed in the context of relatively low-dimensional surface shape models, based on the m-rep formulation [33]. An alternative means of obtaining an FWE combining function permutation test with greatly reduced memory requirements is to combine corrected p-values (instead of correcting combined ones) [28]. In this approach the need to store the complete permutation distribution is avoided, at the computational expense of having to perform a second consideration of the permutation test: a first run is used to obtain the permutation distributions for the maxima of the partial tests; in the second run, comparison of the partial statistics with these permutation distributions yields partial corrected p-values

[10] Hayasaka and Nichols point out that cluster mass is not strictly a consistent combining function, as defined by Pesarin [31], but that it may perform approximately consistently in practice [28]; in fact, in their simulations and real data, it performs very well in most cases.

at every voxel and for each permutation, these are then combined, and the maximum of the combining function over the voxels can be recorded for each permutation; comparison of the original permutation's combined corrected p-value statistic with this distribution gives the overall corrected p-value at each voxel.

It is also possible to combine parametric p-values, or raw t-values at each voxel, which avoids the need for storing the complete permutation distribution; only the image-wise maximum of the combined values need be stored for each permutation in order to determine permutation-based FWE p-values. This approach was taken by Hayasaka et al. [34] in their analysis of grey-matter density and perfusion in Alzheimer's Disease. Terriberry et al. [33] state that the disadvantage of this method is that it does not avoid problems of differing scale or of strong correlation between the multivariate components. The issue of scaling is clear, but it is not obvious why Terriberry et al. feel the test will handle correlation differently when not based on p-values. In contrast, Hayasaka and Nichols postulated:

> We suspect that there is very little effect on the sensitivity and the specificity of the test with our use of partial P values, compared to using the actual peak intensity and cluster size information directly

and when using direct combination of raw statistics, Hayasaka et al. [34] still stated 'the cross-modality correlation is implicitly accounted for'.

Multivariate combining function permutation tests will not be considered further here, but note that since the procedure essentially involves a combination of the results of some standard permutation tests, the work on permutation testing for general linear models below is of direct utility in the combining function setting. We are unaware of any studies in an image analysis context that compare the two alternatives for combining-function permutation-based FWE control discussed above, or that compare the use of p-values to raw statistics in combining functions; these questions could be a useful topic of future work.

2.4 Permutation testing for general linear models

2.4.1 Exact cases

In section 2.2 we presented two simple examples for which the permutation test is exact, simple regression and a two-sample t-test. The logic motivating the two-sample t-test generalises to one-way ANOVA — under the null hypothesis, the observations should be exchangeable between the levels, allowing an exact test of the main effect. Combinatorics shows that

$$\frac{(\sum_{l=1}^{L} n_l)!}{\prod_{l=1}^{L}(n_l!)} \tag{2.2}$$

permutations are possible with n_l observations in level l, where as noted already, random sampling from this set preserves the exactness of the test.

A closely related argument (made as early as 1935 by Fisher [35], and described in some detail by [1]) can be made that for a paired two-sample t-test, the order of each pair

may be randomised under the null-hypothesis, or equivalently, that the signs of the paired differences can be randomly flipped, giving an exact test with 2^n possible 'permutations'.[11] It seems natural to extend this to the case of a one-sample t-test, however, as pointed out by Manly [1] the justification is slightly weakened, and consequently, the required assumptions grow to include that the distribution of the observations is symmetric.

The one-way ANOVA situation can be generalised to tests of main effects in multi-way ANOVA designs if the set of permutations is restricted such that they mix the levels of the factor in question, while preserving the levels of the other factor(s). Examples can be found in [12, 36, 37], and discussion of the theoretical concepts of weak and partial exchangeability in [6]. This provides an exact test, but with reduced power as the number of usable permutations decreases [36]. If a discrete nuisance-covariate is present, a similar strategy can be used, permuting the measures within equal values of the covariate, where the set of permutations reduces to the identity in the limit as the number of non-unique values of the covariate tends to zero.

The above models have exact permutation tests under fairly general assumptions, in particular, they pose no obvious difficulties for observational studies. If one can make additional assumptions thanks to the nature of a designed experiment, then arguments can be made that more general models also have exact permutation tests. For example if all of the (interest and nuisance) covariates X are randomly assigned, before measurements Y are made, then randomisation of the observed measurements (referred to below as the strategy 'Shuffle-Y') provides exact tests for multiple and partial correlations [1]. Alternatively, if one or more interest covariates Z are randomly assigned, before measurements are made for the data Y and for one or more nuisance-covariates X_0, then a permutation test using randomisation of the interest covariates ('Shuffle-Z') while keeping the data and nuisance covariates matched can be argued to be exact [38]. Further arguments can be found in Manly [1] pp.180–181; and counter-arguments (though more aimed at the practical misapplication of Shuffle-Y and Shuffle-Z than their theoretical justification) may be found in [9, 39].

2.4.2 Approximate permutation tests for arbitrary designs

In more general cases, including multiple regression in the observational setting, the above exact permutation tests are not applicable. To take a simple example, if the data, an interest covariate and a nuisance-covariate might all be interrelated, an exact test cannot be achieved by permuting the data (as its conditioning on the nuisance-covariate will not be preserved), and nor is Shuffle-Z exact, since the changing relationship between the interest covariates and the confounds will alter the statistics. Intuitively, in many such situations where the data are not themselves exchangeable, one could argue that the errors in the underlying regression model would be exchangeable. As a specific example, a general linear model with non-Gaussian but IID (or compound symmetric) errors would satisfy

[11] Technically, these are not permutations. In analogy with the definition of a shuffling matrix in 2.2.1, these sign-flippings could be performed with a diagonal flipping matrix F with $f_{ii} = \pm 1$, satisfying $F = F^T = F^{-1}$.

this assumption. We now consider a hypothetical exact test in such situations, followed by approximations towards it.

For the remainder of this section we consider the following (possibly multivariate) general linear model:

$$\begin{aligned} Y &= XB + \mathcal{E} \\ &= \left[\, X_1 \,\vdots\, X_0 \,\right] \begin{pmatrix} B_1 \\ \cdots \\ B_0 \end{pmatrix} + \mathcal{E} \end{aligned}$$

where we have partitioned the design matrix into the effects of interest X_1 and the nuisance-covariates X_0. We shall also refer to the interest covariates as Z (for example in the algorithm 'Shuffle-Z'), as this is common in the literature, and should cause no confusion here.[12] In appendix A.4.8, it is shown that any estimable contrast in any general linear models may be re-written in this form, i.e. with the new contrast being a 1 (or identity matrix) over the column(s) of interest with zeros over the nuisance columns.

Importantly, it is not required here that the error $\mathcal{E}$ be IID Gaussian, but it will be assumed to be exchangeable.[13] It is obvious, but perhaps helpful, to point out, that under the null hypothesis, one has the reduced model $Y = X_0 B_r + \mathcal{E}$, with the same exchangeable errors $\mathcal{E}$. However, the full and reduced models do not have the same estimated residuals: $E = RY \neq E_0 = R_0 Y$.

Anderson and Robinson [40] seem to have been the first authors to observe that a hypothetically exact permutation test would require knowledge of the true (unobservable) B or B_r in order to recover the unobservable exchangeable errors $\mathcal{E}$. In particular, under a single-interest, single-nuisance model, with a test statistic of squared partial correlation, they consider a test which can be seen to be a special case of the following strategy:

- Remove the true nuisance from the data, to recover the errors (under the null hypothesis)
- permute these exchangeable errors
- add the true nuisance back to the permuted errors, to produce the equivalent of new permuted data (not actually a permutation of Y)
- use the new data in the original regression model, testing B_1 with adjustment for the effect of B_0 (now using the estimated, rather than true nuisance effect, for consistency with the original test statistic).

In other words, testing the new data

$$Y_{AR}^{S} = X_0 B_r + S(Y - X_0 B_r). \tag{2.3}$$

[12] Note though that the literature commonly denotes the nuisance as X, which we have avoided since it would cause confusion with the rest of this chapter.

[13] In the multivariate case, 'exchanging' Y or $\mathcal{E}$ refers to permutation of the rows of the matrix, as will be performed by pre-multiplication with a permutation matrix S.

Anderson and Robinson [40] then argue that an obvious approximation to the above exact but unrealisable strategy is to instead adjust for the least squares estimated rather than true nuisance effect, and therefore to test

$$\begin{aligned} Y_{FL}^{S} &= X_0\hat{B}_r + S(Y - X_0\hat{B}_r), \\ &= P_0 Y + SR_0 Y. \end{aligned} \tag{2.4}$$

Freedman and Lane [41] introduced this concept, though not directly in a permutation testing framework. Still and White [42] used what is essentially a special case of the idea in the context of testing for an interaction in ANOVA. For example, an often quoted expression [36, 37, 43] for testing an interaction in a two-way ANOVA model is

$$y^*_{ijk} = y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot}$$

where the dots denote averaging over subscripts. This expression is equivalent to the arguably simpler $R_0 y$, for a reduced design matrix X_0 modelling the two main effects but not their interaction. Note though, despite the apparently greater generality and simplicity of $R_0 y$, the above papers consider more complex ANOVA designs for which the standard Freedman-Lane method is not directly applicable, for example with nested factors [36] or with one or both factors involving repeated measures [37].

It can be shown that the Freedman-Lane approximation has an expected asymptotic correlation of 1 with the hypothetical exact test [40]. Interestingly, despite this apparently strong theoretical foundation, Welch [44, 697] has objected to the Freedman-Lane strategy on the following grounds:

> permuting the residuals is seen to consider all experiments where [the interest covariate(s)] remains constant but the response and rows of [the nuisance covariate(s)] are permuted. Thus [the complete design matrix] is not constant and the ancillarity principle is violated.

This passage is unfortunately rather brief, and not entirely clear; my interpretation of it is that it is the initial regression on only part of the design matrix that violates ancillarity. However, it also seems possible that Welch is mistaken in his interpretation of Freedman and Lane's method. This view is indirectly supported by the following passage from Anderson and Legendre [8], in which they demonstrate an awareness of Welch's point, but appear to believe that it is only Shuffle-Z, and not Freedman-Lane, which violates ancillarity:

> We restricted our attention to methods which do not ignore a potential relationship (collinearity) between the predictor variables (i.e. methods which do not disobey the principle of ancillarity, which means relatedness: Welch, 1990; ter Braak, 1992). This limits the discussion to methods which permute either the raw data values (Y) or residuals of some kind, as opposed to permuting predictor variables.

Their citations are to our [44] and [45] respectively. Welch's suggested method for permutation testing is to identify a 'maximal invariant' with respect to nuisance-parameter location shifts, then to find minimal sufficient statistics for the invariance-reduced data, and then to construct a permutation test relying on these statistics [44]. Kennedy suggests that the residualisation procedure of Freedman-Lane can be viewed as creating such a maximal invariant [39] (in fact, Kennedy says 'it seems to be a better maximal invariant'). It seems that further mathematical research is needed to clarify these apparently contradictory viewpoints.

We note that although the interpretation is eased by considering a process of removal and re-addition of the nuisance effects with the permutation step sandwiched in the middle, the re-addition is actually unnecessary if the new data are regressed on the complete design, since the added nuisance will simply change $\hat{B}_0$ and not the estimated parameters of interest or sum-squared error, as discussed in appendix A.4.8. Hence, instead of Y_{FL}^S from (2.4), the Freedman-Lane method can regress $E_0^S = SR_0Y$ against $[X_1\ X_0]$. As already noted, permutation of the data by S is equivalent to permutation of the design by S^T, so we may alternatively regress R_0Y against $S^T[X_1\ X_0]$ with the same contrast.

Freedman-Lane permutes the reduced-model residuals, $E_0 = R_0Y$; we could also consider the residuals of the full model $E = RY$. This forms the basis of a permutation strategy suggested by ter Braak [45],[14] which he argues should have greater power. The rationale is that subtracting the fitted interest-covariate model before the permutation procedure should reduce the variance of the estimated parameters of interest under permutation. This is considered further by Anderson and Legendre [8, pp. 280–281, figs. 1 & 6]. Note that the optional re-addition of the nuisance in Freedman-Lane, discussed in the preceding paragraph, has even greater relevance to ter Braak's method: if the full-model fitted data PY is added back to the permuted residuals, the original estimated interest X_1B_1 is part of the permuted data, and therefore the statistic should test the estimated interest under permuted B_1^S against the original B_1, instead of testing against zero, as with the other methods. Anderson and Legendre [8] pointed this out, and noted that the procedure is simplified if the fitted model is not restored after permuting the full-model residuals. The ter Braak and Freedman-Lane methods can be respectively termed permutation under the full model and permutation under the reduced model, in which case Shuffle-Y can be seen in the same framework as permutation of the 'residuals' under a null model [1]. Anderson and Robinson [40] showed that all three of these methods converge asymptotically to the same (standard normal) distribution, in terms of the statistic $\sqrt{n}\rho$, though their values may differ for each permutation.

2.4.3 Nuisance-orthogonalisation and related methods

We now discuss three further permutation testing strategies. It is helpful to consider the alternative regression formulations from appendix A.4.10. The Freedman-Lane approach can equivalently perform any of the following regressions (using the notation

[14] He also discuss a bootstrap methodology using the same full-model residuals.

data:design:contrast)

$$E_0^S : X : C, \tag{2.5}$$

$$E_0^S : [X_1\ X_0] : C_p, \tag{2.6}$$

$$R_0 E_0^S : R_0 X_1 : I_{r_1}, \tag{2.7}$$

$$R_0 S R_0 Y : R_0 X_1 : I_{r_1}. \tag{2.8}$$

Notice from the last of these expressions that this form of the Freedman-Lane method re-orthogonalises the permuted orthogonalised data. The need for this can be understood by noting that orthogonalisation does not make R_0Y and X_0 completely independent; it only linearly decorrelates them. After permutation SR_0Y can exhibit some new correlation with X_0, by regressing against the complete model with $[X_1\ X_0]$ this new correlation is adjusted for in the inference on B_1, hence if we wish to regress just on R_0X_1 we must explicitly adjust for this correlation with a second orthogonalisation with respect to X_0. An equivalent explanation of this, is to note that Freedman-Lane is attempting to approximate the unobservable true nuisance [40], and its estimates of this nuisance can be different for different permutations [8].

However, if we consider just the estimated interest-parameters from (2.8), we have:

$$\begin{aligned} B_1 &= (X_1^T R_0 R_0 X_1)^+ X_1^T R_0 R_0 S R_0 Y \\ &= (X_1^T R_0 X_1)^+ X_1^T R_0 S R_0 Y \end{aligned}$$

and we note that the second of these expressions would be consistent with the regression $SR_0Y : R_0X_1 : I_{r_1}$, (i.e. without the post-permutation orthogonalisation). This observation of equivalent interest-parameter estimates wrongly led Kennedy to believe that this simpler regression was completely equivalent to the Freedman-Lane method, and to propose this form as a permutation testing strategy [9, 39].

Anderson and Robinson [40] demonstrated that in terms of partial regression coefficients in a single-interest single-nuisance model,

$$\rho_{FL}^2 = \frac{\rho_K^2}{1 - A_S^2},$$

where A_S^2 is the (non-negative) squared correlation coefficient between X_0 and SR_0y. Hence, the value of ρ_K^2 for the original unpermuted data (for which $A_S^2 = 0$) which equals ρ_{FL}^2 will seem larger in comparison with the permuted values $\rho_K^2 = (1 - A_S^2)\rho_{FL}^2$ than under the Freedman-Lane method, and therefore Kennedy's method may be expected to be less conservative than Freedman-Lane. The difference disappears asymptotically [40], but in practice, several Monte Carlo evaluations have found Kennedy's method to be anti-conservative [1, 8, 40].

Further relationship of Freedman-Lane to Kennedy's method

We now extend Anderson and Robinson's result to more general models. First, note that the equivalence of the interest-parameter estimates ensures that

$$SS_H = \hat{B}^T C(C^T(X^T X)^+ C)^+ C^T \hat{B}$$

is equivalent for Kennedy and Freedman-Lane. For SS_E, using the expression corresponding to regression model (2.8), with (FL) and without (K) the re-orthogonalisation following permutation, Kennedy's estimate is larger by:

$$\begin{aligned} SS_E^K - SS_E^{FL} &= Y^T R_0 S^T R_1^* S R_0 Y \\ &\quad - Y^T R_0 S^T R_0 R_1^* R_0 S R_0 Y \\ &= Y^T R_0 S^T (R_1^* - R_0 R_1^* R_0) S R_0 Y. \end{aligned} \tag{2.9}$$

In appendix A.4.9 we showed that the regression model with both data and interest-covariates orthogonalised with respect to the nuisance no longer needed the nuisance-covariates. Here, it will be helpful to consider an unusual model in which the interest-covariates have been orthogonalised and the nuisance-covariates dropped, but without orthogonalising the data. Such a model will have altered SS_E and SS_R and hence different statistics, so may appear uninteresting. However, in appendix A.4.7, we showed that SS_H can be represented in terms of $R_0 XC = R_0 X1 = X_1^*$ alone, meaning that although SS_E and SS_R change, they do so in a way that preserves SS_H. Since this new model has only X_1^*, its reduced model is empty, with an identity residual forming matrix. Writing its full model residual forming matrix as R_1^*, we have

$$\begin{aligned} SS_H &= Y^T R_0 Y - Y^T R Y \\ &= Y^T Y - Y^T R_1^* Y, \end{aligned} \tag{2.10}$$

and the arbitrariness of Y therefore implies

$$R_0 - R = I - R_1^*. \tag{2.11}$$

If we now consider the effect of orthogonalising the data, replacing Y by $R_0 Y$ in (2.10) gives

$$\begin{aligned} SS_H &= Y^T R_0 Y - Y^T R_0 R_1^* R_0 Y \\ SS_E &= Y^T R Y = Y^T R_0 Y - SS_H \\ &= Y^T R_0 R_1^* R_0 Y \\ &\Rightarrow R_0 R_1^* R_0 = R. \end{aligned} \tag{2.12}$$

Using equations (2.12) and (2.11)

$$\begin{aligned} R_1^* - R_0 R_1^* R_0 &= R_1^* - R \\ &= I - R_0 = P_0, \end{aligned}$$

and so the discrepancy in (2.9) is

$$SS_E^K - SS_E^{FL} = Y^T R_0 S^T(P_0) S R_0 Y,$$

For a univariate model,[15] this can be written

$$SS_E^K - SS_E^{FL} = \|SR_0 y\|_{P_0}^2,$$

where P_0 is a positive semi-definite projection matrix, implying that Kennedy's method will estimate greater (or equal) SS_E in each permutation compared to Freedman-Lane's method, producing smaller statistics, and hence being more likely to reject the null hypothesis when comparing the original (equal) statistic to the smaller permutation distribution.

Interestingly, given that Kennedy and Cade emphasise the need for pivotal statistics in permutation tests, we note that in the case of a single interest-covariate and univariate data, Kennedy's method reduces to a simple-regression of $SR_0 y$ against $X_1^* = R_0 X_1$ (without a constant term), and therefore, as we showed in section 2.2.2, a permutation test using the pivotal t-statistic is equivalent to one using $\hat{b}_1$ with Kennedy's method.

Finally, we note that the equivalence of $\hat{b}_1$ under Kennedy's method and Freedman and Lane's, together with the empirically poor performance of Kennedy's method (regardless of whether it uses t or $\hat{b}_1$), discredits the use of Freedman-Lane with $\hat{b}_1$, supporting Kennedy's argument in favour of pivotal statistics in approximate permutation tests.[16] We return to this point below, in sections 2.4.4 and 2.6.1.

A second permutation strategy was considered in [39] (attributed to Levin and Robbins [46]) and evaluated by Kennedy and Cade [9]. The 'Residualise-Y' or Adjust-Y method seems to be based on the facts that (i) orthogonalising the data with respect to the nuisance doesn't affect the interest-parameter estimates, and (ii) if both data and interest-covarites are orthogonalised with R_0 then X_0 can be dropped from the partitioned model, as discussed in appendix A.4.9. However, for reasons that are unclear, the Adjust-Y method drops the nuisance-covariates from the regression, without first orthogonalising the interest-covariates. The resulting interest-parameter estimates, $\hat{B}_1^{\text{AY}} = (Z^T Z)^+ Z^T S R_0 Y$, are not equal to the original $\hat{B}_1$ under the identity permutation, as noted by Kennedy. However, there appears to be no requirement for this to be the case for a valid permutation test, however intuitively reasonable the property may appear. Kennedy and Cade [9] found in Monte Carlo studies of size that Adjust-Y exhibited a type I error below the nominal value. It therefore appears a valid test. The same authors argued on (slightly heuristic)

[15] In the multivariate case, the same conclusion seems intuitively likely, though it is more difficult to prove. This is empirically investigated using Monte Carlo simulations in section 2.5.

[16] In contrast to situations where the nature of the design means that the permutation test is exact with 'virtually any statistic' [5].

theoretical grounds that Adjust-Y should have low power, however they did not investigate power in their simulation studies, so (as explained later, in section 2.5) it remains possible that Adjust-Y could perform well in practice.

The third permutation strategy we consider here is an adaptation of Shuffle-Z, based on the equivalence of the alternative regression formulations (A.25) and (A.26):

$$Y : [Z\ X_0] : C_p,$$
$$Y : [R_0 Z\ X_0] : C_p.$$

These alternatives naturally lead to the following two permutation strategies:

$$Y : [S^T Z\ X_0] : C_p,$$
$$Y : [S^T R_0 Z\ X_0] : C_p,$$

where the first is the conventional Shuffle-Z, and the second is a new method, proposed by Steve Smith of FMRIB, Oxford (personal communication with Tom Nichols). The only mention of this method in the literature appears to be the following comment, in a paper by O'Gorman [47], who uses standard Shuffle-Z, and appears not to have since followed up the suggestion:

> A referee suggested permuting the residuals of the regression of Z on X [X_0 here], and then adding them to the predicted values of Z.

As discussed with Freedman-Lane and ter Braak's method earlier, there is actually no need to add the predicted $P_0 Z$ back to the permuted $S^T R_0 Z$, as this only affects the nuisance estimates. To see that Shuffle-Z and Smith's method are not equivalent, note that the estimated interest-parameters after permutation are:

$$\hat{B}_1^{\mathrm{SZ}} = (Z^T S R_0 S^T Z)^+ Z^T S R_0 Y$$
$$\hat{B}_1^{\mathrm{SS}} = (Z^T R_0 S R_0 S^T R_0 Z)^+ Z^T R_0 S R_0 Y$$

which are identical only if $S = I$, in which case they both recover the original least-squares interest-parameters $\hat{B}_1 = (Z^T R_0 Z)^+ Z^T R_0 Y$.

O'Gorman's interest in Shuffle-Z [47] was motivated by the fact that it leaves the data and the nuisance paired, which made it (and Smith's method) applicable to an adaptive test [48]. This adaptive test attempts to reduce the effect of outliers in the data by using weights derived from the residuals of the reduced model, which is unchanged under permutation of the interest, but would be changed under Freedman and Lane's method.

2.4.4 Transformed-residual permutation strategies

Anderson and Robinson's [40] exact method is based on the principle that the errors $\mathcal{E}$ are exchangeable under the standard assumptions. The Freedman-Lane and ter Braak approximations to this exact method replace the unobservable true errors with least-squares residuals E or E_0. As well as being intuitively reasonable, it can be shown [49]

that e is the best linear unbiased estimate of ε. From $Y = XB + \mathcal{E}$, $E = RY = R\mathcal{E}$, we observe $\mathrm{E}[\mathcal{E} - E] = \mathrm{E}[(I - R)\mathcal{E}] = (I - R)\mathrm{E}[\mathcal{E}] = 0$, showing the residuals are unbiased; their linearity in the data is immediately obvious. Under the assumed null hypothesis, E_0 clearly satisfies the same properties. Theil [49] (p.195) shows that the (univariate) residuals are the best linear unbiased estimate in the sense that $\mathrm{V}[e - \varepsilon] = \sigma^2(I - R)$, while $\mathrm{V}[\tilde{e} - \varepsilon]$ corresponding to any other linear unbiased estimate $\tilde{e}$ exceeds the first covariance by a positive semi-definite matrix.[17]

However, the covariance matrix of the residuals is no longer a scalar multiple of the identity matrix:

$$\mathrm{V}[e] = \mathrm{V}[R\varepsilon] = R\mathrm{V}[\varepsilon]R^T = \sigma^2 RR^T = \sigma^2 R, \tag{2.13}$$

which implies that the residuals are not exchangeable.[18] In particular, the rank-deficiency of R, $\mathrm{rank}(R) = n - \mathrm{rank}(X)$, ensures that dependencies exist between the residuals (for example, in the simplest case of $X = 1_{n\times 1}$ the residuals are forced to sum to zero). This is also true of the reduced model residuals used in the Freedman-Lane method.

Huh and Jhun observed this fact, and suggested the residuals should be transformed so as to restore their exchangeability [50]. With reference to appendix A.2.3, the compact singular value decomposition of the residual-forming projection matrix is $R = UU^T$ for an $n \times \mathrm{rank}(R)$ matrix U satisfying $U^TU = I$. If one considers the $\mathrm{rank}(R)$-dimensional vector of transformed residuals $e^* = U^Te = U^TRy$, their covariance matrix is

$$\begin{aligned}\mathrm{V}[e^*] &= \mathrm{V}\left[U^TR\varepsilon\right] = U^TR\mathrm{V}[\varepsilon]R^TU \\ &= \sigma^2U^TRU = \sigma^2U^T(UU^T)U \\ &= \sigma^2(U^TU)(U^TU) = \sigma^2 I,\end{aligned} \tag{2.14}$$

meaning that e^* is exchangeable. The same should be true for the rows of E^*, and for the reduced-model residuals $E_0^* = U_0^TR_0Y$ where $R_0 = U_0U_0^T$.

Due to this restored exchangeability, Huh and Jhun claimed that a permutation test using the transformed reduced-model residuals E_0^* (in a Kennedy-like framework) would be an exact permutation test [50]. This concept will now be explored in some detail, initially considering full-model residuals.

First, note that the expression $E^* = U^TRY$ can be simplified, since

$$\begin{aligned}U^TR &= U^T(UU^T) \\ &= (U^TU)U^T \\ &= U^T = (RU)^T; \\ U_0^TR_0 &= U_0^T = (R_0U_0)^T.\end{aligned}$$

Next, observe that the transformed residuals can be back-transformed to return the original residuals $UE^* = UU^TY = RY = E$.

[17] In the multivariate case, each column of E satisfies this property with respect to each column of $\mathcal{E}$.
[18] In the multivariate case, $\mathrm{V}[E] = RVR$ is not block diagonal, so E does not have exchangeable rows.

Now, consider ter Braak's permutation test, using full-model residuals:

$$\begin{aligned} Y_{tB}^S &= PY + SRY \\ &= PY + SUU^TY. \end{aligned}$$

If the permutation step is applied to the transformed residuals E^*, (using a rank$(R) \times$ rank(R) permutation matrix S^*)[19] before back-transforming them and adding back the fitted model, this becomes:

$$Y_U^S = PY + US^*U^TY.$$

At this point, note that for the original labelling $S^* = I$, the above is exactly equivalent to the original data, so there is no need to show the equivalence of $\hat{B}$, SS_E, etc.[20] Surprisingly, we find that although non-trivial permutation alters the data ($Y_U^S \neq Y$) it does not alter the fitted model. Because U is in the null space of P, $PU = 0$, giving

$$PY_U^S = PY + PUS^*U^TY = PY.$$

Similarly, the parameter estimates are also unaltered because X and U are orthogonal, as can be seen by noting $X = PX$ meaning $X^TU = X^TPU = 0$ and hence

$$\begin{aligned} X^+Y_U^S &= (X^TX)^+X^TY_U^S \\ &= (X^TX)^+X^TPY + (X^TX)^+X^TUS^*U^TY \\ &= (X^TX)^+X^TY = X^+Y. \end{aligned}$$

Finally, note that although the residuals generally do differ, SS_E does not:

$$\begin{aligned} RY_U^S &= RPY + RUS^*U^TY = US^*U^TY \\ (Y_U^S)^TRY_U^S &= Y^TPUS^*U^TY + Y^TU(S^*)^TU^TUS^*U^TY \\ &= Y^TU(S^*)^TS^*U^TY = Y^TUU^TY = Y^TRY. \end{aligned}$$

This ability to derive apparently different data sets with the same estimates was noted in [51], where its relevance to statistics education was emphasised. The invariance of the fit, SS_E and the parameter estimates, implies that the standard statistical tests will also be invariant for Y_U^S, and hence that ter Braak's strategy cannot be used with transformed full-model residuals.

Returning to Huh and Jhun's original proposal [50], where the reduced model residuals are transformed, the new 'permuted' data is given by

$$Y_{HJ}^S = Y_{U_0}^S = P_0Y + U_0S_0U_0^TY,$$

[19] Huh and Jhun [50] argue that the reduced dimensionality of the permutation space (from n to rank$(R) = n -$ rank(X), or $n -$ rank(X_0) for E_0^*) is intuitively sensible, since it precludes tests where DF_R is too low.

[20] We have taken a slightly different path to the original paper; Huh and Jhun initially considered the regression of U_0^TY on $U_0^TX_1$ [50], leading to $\hat{B}_1 = (X_1^TR_0X_1)^+X_1^TR_0Y$, as for (A.27), then they proceed to the back-transformation interpretation; they do not consider the use of full-model residuals.

using a $\text{rank}(R_0) \times \text{rank}(R_0)$ permutation matrix S_0. Jung et al. [43] proposed exactly the same procedure in the context of ANOVA designs. For this new data, only the parameters in a nuisance-only model ($Y = X_0 B_r + \mathcal{E}$) are invariant. The invariance of $\hat{B}_r$ follows from the orthogonality of X_0 and U_0, with a very similar derivation to the full-model invariance. The variability of the interest-parameters under permutation allows the construction of a (non-trivial) permutation distribution.

Although the nuisance-only model has invariant $\hat{B}_r$, the nuisance-parameters $\hat{B}_0$ within the full model are not invariant, because although the orthogonality gives

$$[X_1 \; X_0]^T U_0 = \begin{bmatrix} X_1^T U_0 \\ X_0^T U_0 \end{bmatrix} = \begin{bmatrix} X_1^T U_0 \\ 0_{r_0 \times n} \end{bmatrix},$$

the pseudo-inverse mixes the zero and non-zero rows together. However, if the interest is explicitly orthogonalised with respect to the nuisance, changing the original unpermuted nuisance-parameters, then the new $\hat{B}_0^*$ is invariant. The special case of this result arising for a single interest covariate was shown in Huh and Jhun's discussion of the 'ancillarity' of the nuisance in such a situation [50]. The results presented at the end of section A.4.8 regarding the pseudo-inverse of X_p^* allow us to extend this result to the general (possibly multivariate) model.

One might intuitively expect that Huh and Jhun's use of Kennedy's permutation strategy would share the flaw discussed in section 2.4.3. However, by considering the explicitly orthogonalised version, it is apparent that the problem of Kennedy's method not adjusting for reintroduced correlation between the permuted orthogonalised data and the nuisance will not occur, because of the above-noted invariance of $\hat{B}_0^*$ — orthogonalisation of the data will mean $\hat{B}_0^* = 0$ and remains at zero after permutation. An alternative way of showing this is to consider a Freedman-Lane style permutation strategy using the transformed residuals. First, as with FL, we note that adding back the fitted nuisance is redundant, so we may consider $E_{HJ}^S = U_0 S_0 U_0^T Y$. Now, consider the formulation of FL in (2.7),

$$R_0 E_{FL}^S : R_0 X_1 : I_{r_1}.$$

Replacing E_{FL}^S with E_{HJ}^S, post-permutation re-orthogonalisation has no effect because $R_0 U_0 = U_0$ gives

$$\begin{aligned} R_0 E_{HJ}^S &= R_0 U_0 S_0 U_0^T Y \\ &= U_0 S_0 U_0^T Y = E_{HJ}^S. \end{aligned}$$

Therefore Freedman-Lane and Kennedy style permutation tests are equivalent with Huh and Jhun's transformed residuals. A third interpretation of this fact is possible using a more geometric argument, considering univariate data for simplicity. $R_0 y$ lies in a $\text{rank}(R_0)$-dimensional subspace of $\mathbb{R}^n$, however, permuting it can take it out of this subspace, meaning that a second orthogonalisation is necessary; on the other hand $U^T y$ is a general vector in the reduced space $\mathbb{R}^{\text{rank}(R_0)}$, permutation therefore cannot 'add dimen-

sionality' in the same way.

Interestingly, the suitability of Kennedy's method means that for the Huh-Jhun permutation test, $c^T\hat{b}$ can be used in place of t with equivalent results. Furthermore, since $SS_R = Y^T R_0 Y$ is invariant for Huh-Jhun, SS_E and $SS_H = SS_R - SS_E$ are both permutationally equivalent to F for more general univariate contrasts (recall section 2.2.2).[21]

A Shuffle-Z like variant of Huh-Jhun

We now derive a novel alternative formulation of Huh and Jhun's method, in which only the design matrix needs to be modified, in common with Shuffle-Z and Smith's method. We already have the following equivalent models:

$$\begin{aligned} &E^S_{HJ} : X : C, \\ &E^S_{HJ} : [X_1 \ X_0] : C_p, \\ &E^S_{HJ} : R_0 X_1 : I_{r_1}. \end{aligned}$$

The last of these can be expanded to

$$U_0 S_0 U_0^T Y : R_0 X_1 : I_{r_1}, \tag{2.15}$$

with estimated parameters

$$\begin{aligned} \hat{B}^S_1 &= (X_1^T R_0 X_1)^+ X_1^T R_0 U_0 S_0 U_0^T Y, \\ \hat{B}^S_1 &= (X_1^T R_0 X_1)^+ X_1^T U_0 S_0 U_0^T Y, \end{aligned}$$

and full model sum of squares

$$\begin{aligned} SS^S_E &= Y^T U_0 S_0^T U_0^T R_1^* U_0 S_0 U_0^T Y \\ &= Y^T U_0 S_0^T U_0^T (I - P_1^*) U_0 S_0 U_0^T Y \\ &= Y^T U_0 S_0^T U_0^T U_0 S_0 U_0^T Y \\ &\quad - Y^T U_0 S_0^T U_0^T P_1^* U_0 S_0 U_0^T Y \\ &= Y^T R_0 Y - Y^T U_0 S_0^T U_0^T P_1^* U_0 S_0 U_0^T Y. \end{aligned}$$

Now, if we pre-multiply data and design in model (2.15) by $U_0 S_0^T U_0^T$, noting $S_0^T S_0 = I$, we obtain

$$\begin{aligned} U_0 U_0^T Y &: U_0 S_0^T U_0^T X_1 : I_{r_1}; \\ \hat{B}^\dagger &= (U_0 S_0^T U_0^T X_1)^+ U_0 U_0^T Y \\ &= (X_1^T U_0 S_0 U_0^T U_0 S_0^T U_0^T X_1)^+ X_1^T U_0 S_0 U_0^T U_0 U_0^T Y \\ &= (X_1^T U_0 U_0^T X_1)^+ X_1^T U_0 S_0 U_0^T Y; \\ &= \hat{B}^S_1 \end{aligned}$$

[21] In the multivariate case the invariance of $|SS_R|$ means Wilks' Λ is permutationally equivalent to $|SS_E|$, but $|SS_H| \neq |SS_R| - |SS_E|$, so $|SS_H|$ is not equivalent to Λ.

and

$$\begin{aligned}
SS_E^{\dagger} &= Y^T U_0 U_0^T (I - P^{\dagger}) U_0 U_0^T Y \\
&= Y^T R_0 Y - Y^T U_0 U_0^T P^{\dagger} U_0 U_0^T Y. \\
P^{\dagger} &= (U_0 S_0^T U_0^T X_1)(U_0 S_0^T U_0^T X_1)^+ \\
&= U_0 S_0^T U_0^T X_1 (X_1^T U_0 S_0 U_0^T U_0 S_0^T U_0^T X_1)^+ X_1^T U_0 S_0 U_0^T \\
&= U_0 S_0^T U_0^T R_0 X_1 (X_1^T R_0 X_1)^+ X_1^T R_0 U_0 S_0 U_0^T \\
&= U_0 S_0^T U_0^T P_1^* U_0 S_0 U_0^T; \\
SS_E^{\dagger} &= Y^T R_0 Y - Y^T U_0 U_0^T (U_0 S_0^T U_0^T P_1^* U_0 S_0 U_0^T) U_0 U_0^T Y \\
&= Y^T R_0 Y - Y^T U_0 S_0^T U_0^T P_1^* U_0 S_0 U_0^T Y \\
&= SS_E^S.
\end{aligned}$$

Since this model contains only the interest covariates, invariance of $\hat{B}_1^S$ and SS_E implies complete equivalence.

Finally, note that the above regression formulation can be written

$$R_0 Y : X_1^{\dagger} : I_{r_1}$$

where $X_1^{\dagger} = U_0 S_0^T U_0^T X_1$, and that this (A.27)-model can be written in (A.25)-form as

$$Y : [X_1^{\dagger}\ X_0] : C_p.$$

Where we see firstly that the data has been left unmodified, and secondly that the nuisance is also in its original form. Hence O'Gorman's requirements for the adaptive test are satisfied [47]. Comparing the above to Shuffle-Z and Smith's method, observe that instead of shuffling either Z or $R_0 Z$ with S^T, we have instead modified the original interest-covariates using a rank(R_0)-dimensional permutation S_0^T sandwiched between transformations to and from the reduced permutation space. As for Shuffle-Z and Smith, there are computational benefits from not having to recompute R_0 for each permutation.

It might initially appear that a similar trick could be possible with Freedman and Lane's method, however, this is not the case. Considering the formulation in (2.8), one cannot transfer the permutation from $R_0 S R_0 Y$ to the design, because it is not possible to expose S on the left hand side, since R_0 has no left-inverse (recall from section A.3 that the pseudo-inverse recovers the left-inverse where it exists, but that the pseudo-inverse of a projection matrix is itself). Considering instead the formulation in (2.6), the following manipulation is possible

$$\begin{aligned}
E_0^S &: [X_1\ X_0] : [I_{r_1}\ 0_{r_1 \times p}]^T \\
SR_0 Y &: [X_1\ X_0] : [I_{r_1}\ 0_{r_1 \times p}]^T \\
R_0 Y &: S^T [X_1\ X_0] : [I_{r_1}\ 0_{r_1 \times p}]^T \\
Y &: [S^T X_1\ S^T X_0\ X_0] : [I_{r_1}\ 0_{r_1 \times p}\ 0_{r_1 \times p}]^T,
\end{aligned} \tag{2.16}$$

but leads to a new effective set of nuisance-covariates which must span the combined space of $[S^T X_0 \; X_0]$, hence O'Gorman's requirement is not met, and the computational efficiency is reduced through needing to recompute the new $R_0^S = I - [S^T X_0 \; X_0][S^T X_0 \; X_0]^+$ for each permutation.

An interesting parallel between Freedman-Lane and Shuffle-Y can be drawn by transferring the permutation in equation (2.16) to the data:

$$SY : [X_1 \; X_0 \; SX_0] : [I_{r_1} \; 0_{r_1 \times p} \; 0_{r_1 \times p}]^T. \tag{2.17}$$

In this light, Freedman-Lane extends Shuffle-Y to adjust the permuted data for the permuted (as well as the original) nuisance.

Using a similar derivation, not included here, it is also possible to show that model (2.15) is equivalent to

$$S_0 U_0^T Y : U_0^T X_1 : I_{r_1}. \tag{2.18}$$

This provides an interesting comparison to Kennedy's method, with respect to which it merely replaces R_0 by U_0^T. The key point is that the reduced dimensionality of $U_0^T Y$ prevents S_0 from re-introducing correlation with X_0 (as discussed in section 2.4.3). A second interesting point is that the reduced dimensionality means that if DF_E is computed for the above model, it is automatically reduced, in contrast to Kennedy's method (see the end of appendix section A.4.9).

Best Linear Unbiased Scalar-covariance residuals

At the start of section 2.4.4 it was noted that the ordinary least squares (OLS) residuals e are the best linear unbiased (BLU) estimator for the unobservable true errors ε, but that they have non-scalar covariance matrix. Huh and Jhun's transformed residuals [50] are also linear in the data, and unbiased in the following sense,[22]

$$\begin{aligned} E^* = U^T Y &= U^T (X\hat{B} + E) = U^T E, \\ &= U^T (XB + \mathcal{E}) = U^T \mathcal{E}, \\ \mathrm{E}[U^T \mathcal{E} - U^T E] &= U^T \mathrm{E}[\mathcal{E} - E] = 0. \end{aligned}$$

As shown earlier, these residuals also have covariance equal to a scalar multiple of an identity matrix. Theil therefore referred to such residuals as linear unbiased scalar-covariance (LUS) residuals [49], and showed that such residuals must be based on an $n \times \mathrm{rank}(R)$ matrix satisfying $U^T X = 0$, and $U^T R U = I$.[23] Although these transformed residuals have lower dimensionality, this corresponds to the $\mathrm{rank}(R)$ dimensionality of the subspace of $\mathbb{R}^n$ in which the OLS residuals lie. Furthermore, the norm of the transformed residuals matches that of the OLS ones: for univariate data $\|e^*\|^2 = y^T U U^T y = y^T R y = e^T e = \|e\|^2$. The above matrix U is non-unique; any orthonormal basis for the column-space of

[22] The full-model residuals are considered here for the sole reason of notational simplicity, none of the properties are affected by the choice of X or X_0 to derive the residual forming projection matrix.

[23] $UU^T = R$ is not a requirement, but can be proven [49] (pp.208–209), and similarly $U^T U = I$.

R satisfies the necessary properties.

Theil additionally showed [52] that a set of residuals can be found which were not only LUS estimates, but were the best such estimate in a certain sense. Theil's derivation of these residuals (and a later simplification of the procedure, reproduced here [53]) was based on partitioning a linear model with full column rank X into:

$$\begin{matrix}\text{p rows}\\ \text{n-p rows}\end{matrix}\begin{bmatrix}y_\alpha\\ y_\gamma\end{bmatrix} = \begin{bmatrix}X_\alpha\\ X_\gamma\end{bmatrix} b + \begin{bmatrix}\varepsilon_\alpha\\ \varepsilon_\gamma\end{bmatrix} = \begin{bmatrix}X_\alpha\\ X_\gamma\end{bmatrix} \hat{b} + \begin{bmatrix}e_\alpha\\ e_\gamma\end{bmatrix} \tag{2.19}$$

whereupon the BLUS residuals are given by the $n - p$ vector

$$\hat{\varepsilon}_\gamma = e_\gamma - X_\gamma X_\alpha^{-1} \left(\sum_{h=1}^{H} \frac{d_h}{1 + d_h} g_h g_h^T \right) e_\alpha \tag{2.20}$$

where d_h^2 are the square roots of the H eigenvalues of $X_\alpha (X^T X)^{-1} X_\alpha^T$ which are less than one (out of p total), and g_h are their corresponding eigenvectors.

These residuals have the minimal value of the following optimality criterion

$$\begin{aligned} \mathrm{E}[(\varepsilon_\gamma - \hat{\varepsilon}_\gamma)^T (\varepsilon_\gamma - \hat{\varepsilon}_\gamma)] &= \mathrm{tr}\left(\mathrm{V}\left[\varepsilon_\gamma - \hat{\varepsilon}_\gamma\right]\right) \\ &= 2\sigma^2 \sum_{k=1}^{p} (1 - d_h). \end{aligned} \tag{2.21}$$

The BLUS residuals are unique [49], for a certain partitioning in (2.19), however, this partitioning itself is arbitrary. Depending on the purpose for which the BLUS residuals are required, different partitions may be preferred. If testing for autocorrelation it is logical to choose a block of adjacent rows for the γ partition, for example the last $n - p$ [54]. If testing for heteroskedasticity in a time-series then choosing the middle p rows for α is likely to be most powerful since $\hat{\varepsilon}_\gamma$ will then be estimated for the times furthest apart [49]. For the purpose of permutation testing, there would seem to be no reason to favour any partitioning a priori. Note though, that some partitions may not be valid for the expression (2.20) since it is possible for a full column rank X to produce a rank-deficient sub-matrix X_α (indeed, this is quite likely in categorical ANOVA-like designs). It turns out that the optimum value achieved in (2.21) is dependent on the chosen partitions, which provides one objective way to choose a partitioning for a permutation test based on BLUS residuals: evaluate (2.21) (or similar; see below) for all ${}_nC_p$ possible partitions and select the 'best'. The value of σ is unknown, but immaterial for the comparison of different partitions, which can simply consider the criterion divided by $2\sigma^2$.

In [53], Theil argued that the following stronger optimality criterion was satisfied. Define $\Sigma = \mathrm{V}\left[\hat{\varepsilon}_\gamma - \varepsilon_\gamma\right]$; for any other estimate $\tilde{\varepsilon}_\gamma$ with $\tilde{\Sigma} = \mathrm{V}\left[\tilde{\varepsilon}_\gamma - \varepsilon_\gamma\right]$, $\Delta = \tilde{\Sigma} - \Sigma \geq 0$ (i.e. Δ is positive semi-definite). However, it was later shown [55] that while the ordered eigenvalues satisfied $\lambda_i(\tilde{\Sigma}) \geq \lambda_i(\Sigma)$ for all i, this did not imply the eigenvalues of the difference $\lambda(\tilde{\Sigma} - \Sigma)$ were positive semi-definite; therefore Δ is not necessarily a positive semi-definite matrix.

While Theil showed [49] that $\hat{\varepsilon}_\gamma$ can be represented in the form $U_B^T e = U_B^T y$, the

procedure for determining U_B is far from straightforward. A simpler derivation was given by Chow [56], who defined $X_\delta = X_\gamma X_\alpha^{-1}$ and showed

$$\begin{aligned} U_\gamma^T &= (I + X_\delta X_\delta^T)^{-1/2} \\ U_\alpha^T &= -U_\gamma^T X_\delta \\ U_B^T &= [U_\alpha^T \; U_\gamma^T]. \end{aligned} \tag{2.22}$$

More recently, a derivation with greater generality has been given by Magnus and Sinha [57]. While Theil assumed a partitioning of the rows into two sets, it is possible to determine residuals $U_B^T y$ that have a scalar covariance matrix and are the best linear unbiased estimate of $M^T\varepsilon$ for an $n \times \text{rank}(R)$ matrix M satisfying $\text{rank}(M^T RM) = \text{rank}(R)$. This constraint ensures that the full rank square matrix $M^T RM$ has a unique positive square root. Magnus and Sinha prove in their appendix that

$$U_B = RM(M^T RM)^{-1/2}, \tag{2.23}$$

and that the optimal value of the criterion can be expressed as[24]

$$\mathrm{V}\left[U_B^T Y - M^T\varepsilon\right] = \sigma^2(I + M^T M - U_B^T M - M^T U_B). \tag{2.24}$$

In the special case that M selects certain error components (rows of multivariate E), as assumed in Theil's original partitioning framework, $M^T M = I$. Considering the scalar criterion as suggested by [55], the above then simplifies to

$$\begin{aligned} \sigma^2\text{tr}\left(2I - U_B^T M + M^T U_B\right) &= 2\sigma^2(n-p) - 2\sigma^2\text{tr}\left(M^T U_B\right) \\ &= 2\sigma^2(n-p) - 2\sigma^2\text{tr}\left(M^T RM(M^T RM)^{-1/2}\right) \\ &= 2\sigma^2(n-p) - 2\sigma^2\text{tr}\left((M^T RM)^{1/2}\right), \end{aligned}$$

which appears similar to the sum of square-root eigenvalues that occurs in (2.21). We have empirically verified that these two expressions are equal, although algebraically showing their equivalence is not straightforward. As noted above, the expression in (2.20) assumes X_α is invertible; (2.23) does not require this, nor even that X is full column rank. If one considers adding a dependent column to an existing full column rank X, it is clear that R will not change, and hence nor will the BLUS residuals derived using (2.23).

Here, we show a novel (though straightforward) simplification of the above results, stemming from the fact that $M^T RM = (RM)^T(RM)$ and hence the square roots of its eigenvalues are the singular values of RM. If we consider the compact singular value

[24]The authors of [57] appear to be either unaware of or unconvinced by [55], since they consider the matrix form of the criterion.

decomposition of $RM = U_{RM}DV^T$,[25] we have the optimality criterion as

$$\mathrm{E}[\|U_B^T Y - M^T\varepsilon\|] = \sigma^2 \mathrm{tr}\left(I + M^T M - 2D\right). \tag{2.25}$$

Furthermore, we can simplify the expression for the BLUS-forming matrix:

$$\begin{aligned} U_B &= RM(M^T RM)^{-1/2} \\ &= U_{RM}DV^T(VD^2V^T)^{-1/2} && (2.26)\\ &= U_{RM}DV^T VD^{-1}V^T && (2.27)\\ &= U_{RM}V^T. && (2.28) \end{aligned}$$

This provides a helpful comparison to the standard Huh-Jhun procedure, for which we can choose the non-unique matrix U from the compact singular value decomposition of R; the BLUS residuals instead use the unique $U_B = U_{RM}V^T$ derived from the compact SVD of RM,[26] for a chosen matrix M.

One might actually expect that Huh and Jhun's LUS residuals and Theil's BLUS residuals are so similar that they would result in identical permutation tests. Indeed, they are both in one-to-one correspondence, since either may be transformed back to the OLS residuals and from there into the other.[27] However, we have empirically verified that $Y^S_{HJ} = Y^S_{U_0}$ and $Y^S_{U_{B0}}$ give different results for the interest-parameters and SS_E, and different permutation distributions over a large number of such S. This non-equivalence is closely related to a property discussed by Commenges [7], that transformations could preserve exchangeability and yet lead to different permutation tests. For example, compound symmetric data is exchangeable under a Gaussian assumption, and may also be whitened to i.i.d. data if its covariance matrix is known, surprisingly, the two sets of exchangeable data produce different results.

Having extended Theil's BLUS residuals to more general prediction of $M^T\varepsilon$ using $U_B^T y$, Magnus and Sinha [57] also considered the converse question: given a linear unbiased estimator $U_L^T y = U_L^T e$ producing a scalar covariance matrix $U_L^T R U_L = I$, does U_L give the BLUS estimator of $M_L^T\varepsilon$ for some matrix M_L? Assuming $\mathbf{C}(U_L) \subseteq \mathbf{C}(R)$, they showed that in fact a whole class of M_L exists, satisfying

$$M_L = U_L Q + XT, \tag{2.29}$$

where T is an arbitrary matrix and Q is a symmetric positive definite matrix. This means that any non-unique U in Huh and Jhun's permutation method is actually the unique BLUS residual forming matrix for some combination of the errors $M_L^T\varepsilon$. This suggests

[25] Using U_{RM} and D to avoid confusion with the already-defined U and the permutation matrix S, and noting that RM is assumed to be full column rank, which means $V = V_f$ and so $VV^T = I$ in addition to the usual $V^TV = I$.

[26] The uniqueness of $U_{RM}V^T$ despite the non-uniqueness of its factors U_{RM} and V, is not immediately obvious, but the above derivation reproduces Theil's unique BLUS residuals from the original procedure [49].

[27] For example $E_B = U_B^T E = U_B^T U E^*$ where $E^* = U^T E$. This point was made by Magnus and Sinha in relation to BLUS residuals and recursive residuals (discussed later) [57].

that for permutation testing, the conventional BLUS estimate for some selection of the errors might not be superior to any of the LUS solutions, U. However, the BLUS residuals are easier to interpret for M chosen a priori. A further advantage to using U_B for a specified (or optimally selected) partitioning, is that it is uniquely reproducible, while different algorithms or computer platforms could find different U from the SVD of R.

The BLUS residuals were originally derived for a univariate regression model. However, given the BLUS residual forming matrix U_B, one can just as easily compute $U_B^T Y$ for multivariate data Y. The properties of linearity and unbiasedness are clear. Using the Kronecker tensor product and vectorisation operator [58] the multivariate residuals $\text{vec}(\mathcal{E})$ have covariance matrix $V \otimes I$, or vectorising by rows, $\text{vec}(\mathcal{E}^T)$ has block diagonal covariance $I \otimes V$ as intuitively expected. The vectorised product $\text{vec}(R\mathcal{E})$ has covariance matrix $V \otimes R$, while $\text{vec}(U_B^T \mathcal{E})$ restores this to $V \otimes I$ [59]. The optimality criterion will be satisfied for each column of $\hat{\mathcal{E}}$ with respect to the corresponding column of $\mathcal{E}$ (this follows logically from the fact that the matrix U_B depends only on X (or X_0) and not on the data. BLUS residuals have been used for multivariate data to address the problem of outlier detection [60].

Recursive residuals

Another type of transformed residual has also been proposed in the literature: 'recursive residuals' are defined in terms of the prediction error from models based on the data (and design) available 'before' the current observation. They are therefore most clearly interpretable for situations where a natural ordering of the data exists. Recursive residuals are reviewed by Kianifard and Swallow, who describe them as 'standardised one-step-ahead forecast errors' [61]. They can also be considered within a more general class of 'conditional residuals' [62].

Denoting the sub-vector of univariate data from y_1 to y_i by $y_{\leq i}$ (so that $y_{\leq n} = y$), the corresponding sub-matrix of the design as $X_{\leq i}$, and the i^{th} row of the design matrix as x_i^T, the recursive residuals e_R^i can be obtained as follows:

$$\hat{b}_{\leq i} = X_{\leq i}^{+} y_{\leq i} \qquad i = \text{rank}(X), \cdots, n-1 \tag{2.30}$$

$$v_i^2 = 1 + x_i^T \left(X_{\leq i-1}^T X_{\leq i-1} \right)^{-1} x_i \qquad i = \text{rank}(X)+1, \cdots, n, \tag{2.31}$$

$$e_R^i = \frac{y_i - x_i^T \hat{b}_{\leq i-1}}{v_i} \qquad i = \text{rank}(X)+1, \cdots, n. \tag{2.32}$$

where $\sigma^2 v_i^2$ is the variance of the forecast error $y_i - x_i^T \hat{b}_{\leq i-1}$. Only $n - \text{rank}(X) = \text{rank}(R)$ recursive residuals can be computed (the same as the number of BLUS residuals) because $\text{rank}(X)$ data points are required before the first $\hat{b}_{\leq i}$ can be estimated.

The recursive residuals defined above have the same distribution as the unobservable errors, i.e. with the usual assumptions, they are independently and identically distributed as $\mathcal{N}(0, \sigma^2)$. The vector of recursive residuals can be written as $e_R = U_R^T y$, because each recursive residual in (2.32) depends linearly on the current and past data. Since the recursive residuals are linear in the data, unbiased ($U_R^T X = 0$ is easily observed), and

have a scalar covariance matrix (implying $U_R^T U_R = I$ and $U_R U_R^T = R$), they are within Theil's class of LUS residuals [57]. The equation in [61] for U_R (their $\mathbf{C}$ is our U_R^T) has unfortunately been incorrectly typeset. The correct expression can be observed from (2.32) above; for completeness, we provide the following verified MATLAB code:

```
UR = zeros(n, n-k);
for r = k+1:n
    invprod = pinv(X(1:r-1, :)' * X(1:r-1, :));
    UR(1:r-1, r-k) = - X(1:r-1, :) * invprod * X(r, :)';
    UR(r, r-k) = 1;
    UR(:, r-k) = UR(:, r-k) / sqrt(1 + X(r, :) * invprod * X(r, :)' );
end
```

The expression in (2.32) assumes that X and all its sub-matrices $X_{\leq i}$ are full column rank. Tobing and McGilchrist derived much more complicated expressions [63], using recurrence-relations for efficient updating of the various terms involved, with adjustments included to allow for the ranks of the sub-matrices to change within the recurrence formulae. The same authors also allowed for multivariate data. For the current purpose, the number of permutations (and voxels for imaging data) are much greater than n, so naïve computation using the above algorithm suffices. We have also observed that simply using the pseudo-inverse (`pinv` in the MATLAB code) results in a matrix which satisfies all the basic properties, as would be expected, since the fundamental idea is to predict y_i using $y_{\leq i}$, and this presents no difficulties with rank-deficient designs. Similarly, predicting multivariate data poses no problems, suggesting that the obvious $U_R^T Y$ retains the usual recursive residual interpretation, as well as clearly satisfying the LUS properties.

Comparing recursive residuals to BLUS residuals, note that since recursive residuals are in Theil's LUS class while BLUS are optimal within this class, it might initially seem obvious that BLUS would be superior. However, this is not necessarily true for two main reasons. First, as pointed out in [61] the optimality of (2.21) does not imply that tests based upon BLUS residuals will be more powerful than tests based on other residuals with worse mean squared approximation error. For example, in testing for a change in the true regression model part way through a time series (an example of a 'structural break'), BLUS outperformed recursive residuals for a test known as cusum, while recursive residuals showed higher power with cusum-of-squares [57]. Secondly, as mentioned when comparing BLUS and Huh-Jhun residuals earlier, Magnus and Sinha's result (equation 2.29) implies that the LUS recursive residuals contain the same information as the BLUS residuals, and are in fact the BLUS predictor of some combination of the errors [57]. This is of particular relevance to permutation testing, where the ability to select a particular set of residuals for BLUS estimation seems to have little value; arguably, a more mixed compound of the errors might be preferable. Conversely, the strongest arguments in favour of recursive residuals for applications other than permutation testing tend to be based on their more natural interpretation and more obvious correspondence with the original data [61]. For example, Magnus and Sinha mention the BLUS residuals going 'out of fashion [because] recursive residuals have a more intuitive appeal' and are believed to be preferable for

testing structural breaks [57]. Tobing and McGilchrist point out that if one considers updating the estimated regression parameters as each new observation is included, the update is proportional to the recursive residual [63], which seems like an attractive property for the detection of outliers in time-series data. However, these intuitive advantages aren't necessarily upheld in practice; from their Monte Carlo power studies, Magnus and Sinha conclude that BLUS are preferable to recursive residuals, even when testing for a structural break, and they argue that the preference for recursive residuals in recent literature is unjustified [57]. Furthermore, for permutation testing, much of the time-series based intuition behind recursive residuals is lost. The recursive residuals seem to have been developed further with regard to dependent data, for example [63] considers their use together with REML estimation of the covariance components, however, this is not of particular interest for the permutation-testing application at hand, where computational demands would be too great for such a combination.

One practical difference between recursive residuals and Theil's BLUS residuals, is that when the latter are computed for a matrix M which selects rows, they do not depend on the order of the selected or unselected rows, but only on which rows are selected. Consider a rank(R) permutation matrix S_B, post-multiplication of M by S_B shuffles the columns, which effectively permutes the rows containing ones, without allowing them to be exchanged with the earlier zero-rows. Because S_B is orthogonal, if the the square root and inverse are considered via the eigen-decomposition of $M^TRM = VD^2V^T$, they change only the orthogonal V and so can be taken outside the square root and inverse operations, which affect only D^2. From (2.23),

$$\begin{aligned} U_B^S &= RMS_B(S_B^TM^TRMS_B)^{-1/2} \\ &= RMS_B(S_B^TVD^2V^TS_B)^{-1/2} \\ &= RMS_BS_B^T(VD^2V^T)^{-1/2}S_B \\ &= RM(M^TRM)^{-1/2}S_B = U_BS_B, \end{aligned}$$

showing that the permutation only permutes the obtained residuals without changing their values: $(U_B^S)^TY = S_B^TU_B^TY = S_B^TE_B$. While it is clear from (2.32) that the recursive residuals are dependent both on the choice of rows to exclude and on the order of the retained rows.

A second difference relevant to permutation testing is that the partition for the optimal BLUS residuals can be objectively selected from a given set (possibly from all selections, though obviously not from all general matrices M), e.g. using (2.25), while the recursive residuals have no expression for the 'quality' of their chosen reordering of the rows. With no a priori reason to select a particular order for the recursive residuals, one could either use a random ordering or (perhaps preferable for reproducibility) simply use the original order, with the first rank(X) errors not estimated.

To the best of our knowledge, neither the BLUS nor recursive residuals have been investigated in a Huh-Jhun style permutation test (though Commenges seems to suggest such a strategy with Theil's BLUS residuals [7]), let alone directly compared. Bootstrap

tests have been proposed using recursive residuals [64] and BLUS residuals [65], though we are again unaware of any direct comparisons (Vinod seems to prefer recursive to BLUS, though provides little in the way of evidence [64]). Grenier and Léger specifically considered the multiple regression problem, and performed Monte Carlo comparisons of the BLUS residuals to OLS residuals, and also to 'standardized' and 'studentized' residuals [65]. They concluded that BLUS were as good and sometimes better than the standardized and studentized alternatives. There are two interesting differences between the bootstrap formulation in [65] and the Huh-Jhun permutation test,[28] firstly, instead of sampling (with replacement) rank(R) transformed residuals and back-transforming them to produce n modified residuals, Grenier and Léger directly sample n values from the rank(R) transformed residuals. Obviously this is not feasible within the permutation testing framework, but it could be of interest to compare the two bootstrap approaches in future work. The second difference is that [65] uses centred BLUS residuals. For a model containing a constant term (in $\mathbf{C}(X)$, even if not explicitly), the OLS residuals are zero-mean; somewhat surprisingly, the BLUS residuals are generally not. Grenier and Léger do not explain their preference for centred residuals, and it seems very slightly flawed for two reasons: the bootstrap samples from centred residuals will generally not be zero mean themselves (only permutations, which can occur as special cases of sampling with replacement, would guarantee this); and the centring process itself will colour the decorrelated BLUS residuals with a compound symmetric correlation structure.[29] While the compound symmetry does not violate exchangeability, there seems no reason to expect it to be helpful. In any case, the *back-transformed* permuted transformed residuals will be zero mean, because $X^TU = 0$ implies that the constant term (within the column space of X) is also orthogonal to U and hence to USU^TY. The same will be true for the reduced-model transformed residuals if the constant term remains in the span of X_0.

Finally, note that all these transformed residuals are only exchangeable in the exact sense if they are derived from (exchangeable) normally distributed data, for other distributions, they achieve only second-moment exchangeability [7]. After considering both linear and non-linear transformations towards exchangeability, Commenges states that for correlated non-normal data, exact permutation tests are unavailable [7].

Having presented three differently motivated strategies for creating i.i.d. residuals, it is natural to wonder whether slightly more general transformations to second-moment exchangeability could be superior. Commenges [7] proves that linear transformations from the OLS residuals to second-moment exchangeable residuals Te must be of the form: $T = GQ + M$, where G is an $m \times m$ exchangeable matrix, Q is an $m \times n$ matrix whose rows form an orthonormal basis for any subspace of $\mathbf{N}(X^T)$, M is a matrix whose rows are in $\mathbf{C}(X)$, and m can be either rank(R) or rank(R) $- 1$.[30] In terms of linear transformations from the data of the form U_C^Ty, since $e = Ry$, $MR = 0$, and $QR = Q$ we must have

[28] In fact, Huh and Jhun [50] also presented a bootstrap test using their LUS residuals with back-transformation as discussed here.

[29] Centring can be enacted with the projection matrix $\bar{M} = I - 1_{n\times 1}1_{n\times 1}^+$, which results in the covariance matrix changing from $\sigma^2 I$ to $\sigma^2\bar{M}$.

[30] The larger option of rank(R) is obviously preferable for limited amounts of data, and is consistent with the dimensionality of the BLUS and recursive residuals.

$U_C^T = GQ$, where we observe that the transformed residuals considered here are obvious special cases with $G = I$ but different bases Q. Since the class of matrices G is very limited ($G = aI + b1_n$) it seems unlikely that other linear transformations to exchangeability will be dramatically different.

2.4.5 Summary of permutation strategies

Table 2.1 summarises all of the permutation testing methods that are analysed in the Monte Carlo evaluations in section 2.5. Several variations on the different methods have been commented on elsewhere in this chapter. For completeness, the most important ones are reiterated in the table, after their primary forms. The alternative version of the exact test is listed to emphasise that under a true alternative hypothesis, the exact test still assumes the null holds and therefore assumes that $\mathcal{E} + X_1 B_1$ retains the exchangeability property of $\mathcal{E}$.

We have attempted to give the primary reference for each method;[31] where a second reference is also given, this is typically a more recent one which we have found to be the most useful. Note that the reference for Smith's method [47] only mentions it in passing; this chapter is believed to provide the first detailed consideration of it. A reference is not given for the optimal BLUS implementation of HJ, which we have termed Theil's method, nor for the recursive residual implementation, Re, since these are presented here for the first time.

2.5 Monte Carlo evaluation studies

Several authors have used Monte Carlo simulation to evaluate the performance of alternative permutation testing approaches under different conditions, e.g. [1, 8, 9, 39, 40, 50]. However, these studies have had certain limitations which motivate further investigation.

Each paper has typically only evaluated a subset of the available methods. In particular, Anderson and Legendre's simulations [8] are probably the most thorough in terms of the number of scenarios considered, but they have deliberately excluded several methods including Shuffle-Z on grounds of design ancillarity, as discussed in section 2.4.2; their study also appeared before transformed-residual strategies were proposed [7, 50]. Huh and Jhun [50] mentioned both Kennedy's method and Freedman-Lane, but seem unaware of Anderson and Legendre's demonstration that Ke is invalid, and compare their new approach only to Ke and not to the superior FL method. A later paper compared the Huh-Jhun technique to FL, tB and SY, but only in the context of pure ANOVA designs [43]. The latest edition of Manly's text-book [1] appears to investigate the most complete set of methods, and includes some comparisons of Huh and Jhun's method to FL and SY for regression, however these simulations have other short-comings mentioned below. Naturally, the novel methods of Sm, Th, and Re do not appear in any existing evaluations.

[31] Still and White [42] arguably developed the method commonly known as Freedman-Lane earlier, though in the context of ANOVA instead of general regression. Judging from a citation by Manly [1], the method may have been developed even earlier by Beaton [66], though we have been unable to obtain this conference paper.

Method	Abbr.	Ref.	Data	Interest	Nuisance	Notes
Exact	Ex	[40]	$S(Y - X_0B_0)$ $S(\mathcal{E} + X_1B_1)$	X_1 X_1	X_0 X_0	(a)
Freedman-Lane	FL	[41] [40]	$S(Y - X_0\hat{B}_0)$ SR_0Y R_0SR_0Y	X_1 X_1 R_0X_1	X_0 X_0	cf. Ex cf. AY cf. Ke
ter Braak	tB	[45] [8]	SRY	X_1	X_0	
Huh-Jhun	HJ	[50] [7]	$U_0S_0U_0^TY$ Y $S_0U_0^TY$	X_1 $U_0S_0^TU_0^TX_1$ $U_0^TX_1$	X_0 X_0	(b) cf. SY cf. Ke
Theil	Th		$U_BS_0U_B^TY$	X_1	X_0	(c)
Rec. Res.	Re		$U_RS_0U_R^TY$	X_1	X_0	(d)
Shuffle-Z	SZ	[38] [1]	Y	S^TX_1	X_0	
Smith	Sm	[47]	Y	$S^TR_0X_1$	X_0	
Shuffle-Y	SY	[67] [1]	SY Y	X_1 S^TX_1	X_0 S^TX_0	 cf. SZ
Adjust-Y	AY	[46] [9]	SR_0Y	X_1		(e)
Kennedy	Ke	[39] [40]	SR_0Y	R_0X_1		(f)

(a) Unobservable true B_0 used; $Y - X_0B_0 \neq R_0Y$
(b) U_0 comes from the compact SVD of R_0
(c) U_B is from (2.28) for an optimal selection matrix under (2.25)
(d) U_R gives recursive residuals based on X_0, dropping first r_0 rows, see § 2.4.4
(e) This method does not give the usual $\hat{B}$ for $S = I$
(f) DF_E must be reduced to account for the DF removed by R_0

Table 2.1: The eleven permutation strategies featured in the Monte Carlo evaluations. S and S_0 are permutation matrices; S is $n \times n$ while S_0 is rank(R_0)-dimensional.

All of the above-mentioned studies feature only a univariate dependent variable. Anderson and Robinson state that their results 'can be readily extended to the case of multivariate response' [40] but they present no simulations for such cases. Interestingly, the same paper includes a comment that parametric multivariate tests are typically less robust to non-normality [40], which should heighten interest in evaluating their permutation-based counterparts.

The literature has also focussed on designs with only one interest covariate, and often a single nuisance, for example [9] evaluates only the single-interest, single-nuisance case, while [8, 40] consider multiple nuisance-covariates but only one interest. While the theoretical extension of the permutation test from t- to F-tests over multiple covariates of interest is trivial, it is not a foregone conclusion that the practical performance of the different measures will show the same patterns.

Furthermore, all studies of which we are aware consider either the basic regression situation with continuous interest and nuisance variables (e.g. [8, 9, 39, 40, 50]) or they consider a pure ANOVA scenario without nuisance-covariates (e.g. [37, 40, 43]). These

simulations therefore do not include the following realistic situations with potentially important practical relevance: regression with continuous interest variable(s) but categorical (e.g. gender) nuisance(s); regression with continuous interest and a mixture of continuous and categorical nuisance-covariates; ANCOVA, or regression with categorical interest variable(s) and either continuous or mixed nuisance covariates.

Various special-case simulations have been proposed to highlight problems with particular methods, for example Kennedy and Cade [9] were the first to suggest an interesting scenario under which they expected Manly's Shuffle-Y method [1] to perform poorly: if the nuisance-covariate contains an unusual 'outlying' value, but the data is generated from the fitted nuisance with additive errors so that there is no real outlying measurement, then after the data are shuffled the nuisance no longer explains the unusual value in the data, which therefore becomes a genuine outlier. We refer to this as the presence of a 'pseudo-outlier'. Other extreme cases include severely skewed error distributions such as cubed exponential [8], or designs with very low degrees of freedom. Often, these special cases have been considered in isolation, but their combinations have not always received such thorough attention. Anderson and Legendre have carried out the most exhaustive experiments in this sense [8], though they have not considered e.g. multiple nuisance-covariates including one with a pseudo-outlier.

Another limitation is that many studies have used relatively few random permutations. For example, 999 permutations (plus the original) has been a common choice [8, 9, 40]. Several of Manly's comparisons use just 99 randomisations [1]. As shown in section 2.5.3, 1000 permutations is not really sufficient for accurate permutation test p-values. A related point is the number of simulated data-sets, and whether different random designs were simulated or whether only the errors were randomly sampled and the designs fixed. For example Kennedy and Cade's simulations [9] and some of Manly's [1] appear to have used a single design with respectively 1000 and 10,000 randomly generated sets of errors. It is clear that multiple designs must be considered in order to avoid accidentally unrealistic situations biasing the results. Furthermore, it seems intuitively reasonable that each design should be tested with multiple sets of errors for the same reason, although this approach does not appear to have been used in the literature.

Lastly, note that some studies have evaluated only size, and not power, e.g. [9, 40]. It may seem natural to expect that tests which are closer to exact under the null hypothesis (i.e. only just valid, rather than very conservative) will be more powerful when the null hypothesis is false. However, according to Manly (p.192) [1], a test which has fewer than expected false positives under the null hypothesis is not necessarily less powerful under the alternative hypothesis. It is therefore necessary to separately evaluate power.

Here, we perform a wide-ranging set of Monte Carlo evaluations, attempting to address the limitations described above, and including a larger number of permutation methods than any other single comparison in the literature. Following the main study into the different GLM permutation methods, two minor experiments are performed to investigate related aspects. Firstly, results from Anderson and Robinson [40] regarding correlations among the statistics for some of the permutation methods are extended to a greater number

of methods and simulations. Secondly, investigations are carried out into two particular questions regarding the number and type of permutations sampled.

2.5.1 Linear model permutation techniques

Experimental setup

Data is generated from the following linear model:

$$Y = X_1 B_1 + X_0 B_0 + \mathcal{E}. \tag{2.33}$$

The data is $n \times m$, and there are r_1 interest and r_0 nuisance covariates. There is no loss of generality in having separate interest and nuisance partitions, compared to the apparently more general case of an arbitrary design with an estimable contrast implicitly defining the interest and nuisance spaces. This is explained in appendix A.4.8, which shows that an equivalent partitioned form can be found for any estimable contrast.

Note that this generative model means that the true error $\mathcal{E}$ is available, along with the true values of the interest and nuisance-parameters B_1 and B_0. The rows of $\mathcal{E}$ (elements of ε in the univariate case) are independently and identically sampled, which ensures that they are genuinely exchangeable. Three different error distributions are considered: the standard normal, for which the parametric test should be optimal; a heavier-tailed t-distribution on 5 degrees of freedom; a severely non-Gaussian highly skewed distribution, raising standard mono-exponential samples to the third power. To evaluate the test size for the various permutation strategies, the data are generated under a true null hypothesis with $B_1 = 0_{r_1 \times m}$, while to evaluate power the alternative hypothesis is true, with $B_1 = \delta 1_{r_1 \times m}$ for scalar effect size δ. We decided to consider the following ranges of non-zero effect sizes: $\delta \in \{2, 4, 6, 8\}$. Although chosen a priori, the results a posteriori reveal a wide range of powers from 4% to 97% being observed for the exact method, endorsing the suitability of this choice.

The use of multivariate data and/or multiple interest covariates necessitate the use of a two-sided test. For simplicity, a two-sided test is therefore also used in the univariate and single-interest cases. The permutation testing methods are compared to parametric results from the normal-theory F-test (or Rao's F approximation in the multivariate case — appendix A.4.4) which is abbreviated as PF in the tables and figures.

The most thorough Monte Carlo evaluations in the permutation-testing literature seem to have used $N_s = 10000$ sets of simulated data [1, 8]. While some of these studies have used single designs with 10,000 noise realisations [1], as argued above, we maintain that multiple designs are required, and believe that using multiple error samples for each design is preferable to ensure that all designs are evaluated reasonably well. We therefore elect to use a total of 10,000 simulations, arising from 100 error realisations being simulated for each of 100 randomly generated designs. Note that the same random design and error realisations are used for evaluating each permutation method.[32] In contrast, an attempt

[32]The same random permutations are also employed within each method (excluding the reduced-space permutation methods) with S or S^T used for the data or design as appropriate.

to compare different permutation methods evaluated for different designs and noise (e.g. performing a meta-study of different papers) would be somewhat less valid due to the random variation in designs and noise.[33]

As mentioned above, there is practical interest in determining the performance of the different permutation testing methods in situations other than purely continuous regression models or purely categorical ANOVA designs. A total of six types of design are considered here, listed in table 2.2. Mixed interest seems of much lower practical relevance, and has therefore not been included in the simulations. It should be noted that although cases DM and DD are similar in form to ANCOVA and ANOVA, we have not used standard factorial designs, nor considered interactions. The disadvantage of this is slight reductions in realism and relevance, but the advantage is that a greater number of random designs are available, which one might hope would include some challenging/pathological examples by chance, thus stressing the methods more acutely.

Code	Interest	Nuisance	Example/Note
CC	Continuous	Continuous	E.g. Multiple regression
DC	Discrete	Continuous	E.g. Two-sample t-test with age as nuisance
CM	Continuous	Mixed	N.B. mixed only possible with $r_0 \geq 2$
DM	Discrete	Mixed	E.g. Two-way ANCOVA
CD	Continuous	Discrete	E.g. Regression, with gender as nuisance
DD	Discrete	Discrete	E.g. Two-way ANOVA

Table 2.2: The six design matrix scenarios considered in the Monte Carlo evaluation.

We argue that the most extreme case of a categorical variable is a Boolean or binary coding variable (for example indicating membership of one of two groups) and that a more general discrete variable with more levels tends towards the case of a continuous variable. For this reason, our discrete (and mixed) designs use binary covariates simulated from `round(rand(n,1))`. The continuous covariates are simply drawn from the standard uniform distribution using `rand(n,1)`.

To generate continuous nuisance-covariate(s) correlated with the (continuous or discrete) interest, we simply add the interest covariate (or the mean of multiple interest covariates) to a new `rand(n,r_0)` matrix. This results in a relatively high expected correlation of about 0.7.[34] Correlations among multiple nuisance-covariates themselves are irrelevant because all the permutation methods keep the rows of multiple nuisances together, and only the space spanned by multiple nuisance columns (and their relation to the interest) affects the statistics. Discrete nuisance-covariates (including those in the mixed nuisance cases) are not made to be correlated with the interest in this way, but are simply generated from `round(rand(n,1))` as the interest were.

For continuous- and mixed-nuisance designs, we add a severe outlier to the first (or only) continuous nuisance-covariate, by replacing one of its values with the value of its

[33] In fact, while some of the error distributions are commonly chosen (most notably, standard normal and cubed-exponential), no standardised method of generating random designs has appeared in the literature, making meta-studies largely impossible.

[34] For two independent standard uniform vectors u_1 and u_2, $\mathrm{V}[u_1] = 1/12$, $\mathrm{V}[u_1 + u_2] = 1/6$ and $\mathrm{V}[u_1, u_1 + u_2] = 1/12$. Giving the correlation of u_1 and $u_1 + u_2$ as $(1/12)/\sqrt{(1/12)(1/6)} = 1/\sqrt{2} \approx 0.7071$.

mean plus ten times its original standard deviation. As discussed earlier, this 'pseudo-outlier' is part of the data-generation process, so there is no true outlier in the regression, as long as the data and the nuisance covariates remain aligned. This was originally proposed to challenge the Shuffle-Y method (which will effectively create real outliers in the permuted data) [9], but will also affect methods based on the residuals, like FL, tB and HJ, among others.

A potential problem when randomly generating many designs (particularly for the discrete or binary cases), is that one or more of the interest columns could lie in the space of the others or the nuisance, and hence be inestimable. Similarly, rank-deficiency within the nuisance-covariates would change the space in which the transformed-residual strategies perform their permutation. To avoid these issues, we repeatedly generate the interest covariates until they are full-rank and non-constant. Then the nuisance covariates are repeatedly sampled until the entire design matrix of interest, constant and non-constant nuisance has full column rank. There is no loss of generality in having a full-rank design (which implies that the separate interest and nuisance partitions will also be full-rank). This can be observed from the fact that the regression equations depend only on the spaces spanned by the full and reduced models (via the projection matrices R and R_0), which are unaffected by the removal of dependent columns.

The design-generation procedure described thus far is independent of n, m or the number of interest or nuisance-covariates (r_1 and r_0 respectively). However, in practice, limits of computational time, and the requirement for relatively straightforward interpretation of the results, restricts the number of cases which can be considered. We chose to perform thorough univariate analyses in a particularly challenging small-sample situation, with $n = 6$, where it is not practical to further lower the degrees of freedom by having more than a single interest and a single nuisance-covariate (plus a constant term). Hence the mixed designs cannot be considered, leaving a total of four designs crossed with three error distributions, for 12 simulations in total. For $n = 6$, we evaluate the exhaustive set of $n! = 720$ permutations. Note that even fewer permutations will be available for designs with discrete interest and discrete nuisance; potentially as few as 30 in the case that both the interest and nuisance consist of a single 1 (in different positions) among zeros.

The smallest n which allows 2 nuisance-covariates in addition to the constant while maintaining the $(n-3)!$ number of available permutations for the transformed-residual methods at a reasonable level is $n = 9$, which gives 720 reduced space permutations. 5040 permutations are available for the other methods (though again, potentially fewer for discrete designs). Arguments could be made for randomly sampling 720 of these, for a 'fair' comparison to the transformed-residual strategies. However, we believe that a realistic comparison must accept that the residual-transformation reduces the available permutations, hence for $n \geq 9$ we randomly sample 5000 permutations, as this number is relatively typical of common practice in neuroimaging. We evaluate $n = 9$ with $r_1 = 2$, $r_0 = 3$ (including the constant nuisance). Univariate and multivariate ($m = 2$) data are considered, for all six design cases, giving 12 scenarios per type of error. Only normal errors are considered here, to maintain a manageable number of simulations.

At higher sample sizes (perhaps more correctly, higher DF_E, since large samples with almost equally large numbers of covariates are still challenging) the performance of the different methods has previously been shown to converge under reasonable circumstances. For this reason, and to partially address the limited number of error distributions used for $n = 9$, we chose to evaluate $n = 16$ with $m = 1$, $r_1 = 2$, $r_0 = 3$, and cubed-exponential errors. Isolated partial simulations (e.g. with only the null hypothesis, or with lower N_s) have also been performed for some other cases, for example $n = 20$ and different numbers of interest and nuisance-covariates. However, no interesting patterns were observed beyond those which apparent from the the situations described thus far, so these partial simulations are not reported in the tables of results.

Performance metrics for evaluation

The fundamental characteristics of a classical statistical test are its size and power, or equivalently, its probability of a false positive and the complement of its probability of a false negative. These measures have been the most common performance metrics for evaluating permutation tests. Typically, the size α at a particular a priori significance level α_0 is reported, often along with a confidence interval based on the number of simulations (see table 2.13 in section 2.5.3), and such results are also presented below. However, from preliminary experiments, it became clear that the performance of the different permutation testing techniques (and the parametric F-test) can change for different values of the arbitrary level α_0. In particular, a plot of the observed α against the expected value over a range of α' is expected to be the straight $y = x$ line under the null hypothesis, but the different methods tend to have rejection rate curves which differ not only in their overall slope or intercept with respect to this line, but also in their variability around it. For this reason, we additionally present tables and figures of several summary measures. Bias in accuracy is summarised by the mean of the error $\alpha - \alpha'$ over a range of α' equal-or-more-significant than the predefined significance level(s), i.e. from one or more values of α_0 down to zero. Variability in accuracy is summarised by the Root-Mean-Square (RMS) of the above error over the same range(s). For simplicity, in the main experiment we focus solely on $\alpha_0 = 0.05$, which seems to be the most common value in neuroimaging, however, we will briefly investigate the impact of different α_0 within the secondary experiment of section 2.5.4.

Under a true null hypothesis, the p-values are expected to have a standard uniform distribution — i.e. one expects 5% of them to be below 0.05, 10% of them to be above 0.9, etc. We therefore present a third performance metric for accuracy under the null hypothesis, based on the Kolmogorov-Smirnov test statistic. The K-S test is one of the most commonly used procedures for testing the hypothesis that a sample comes from a specific distribution; its test statistic is simply the greatest distance between the empirical cumulative distribution function of the sample, $\hat{F}(x)$, and the theoretical cumulative distribution function, $F(x)$, for the assumed distribution: $k = \max_x |F(X) - F_0(X)|$. Any subset of p-values less than or equal to α' will have a uniform distribution over the interval $(0,\ \alpha')$, and hence a linear CDF $F(x) = x/\alpha'$ over this interval. Therefore, a K-S statistic

can be computed over the range(s) of most relevant α' from zero to one or more levels α_0.

Power introduces two major complications with respect to size. Firstly, while the null hypothesis specifies a unique value of the interest-parameters (assumed to be zero here), the alternative hypothesis is true for infinitely many values. As described in section 2.5.1, a representative set of effect sizes can be used for simulation, such that a broad range of powers is observed. However, for ease of interpretation of the results, we usually summarise these by averaging together the observed rejection rates α,[35] As for size, we further summarise by averaging over a range of α' from zero to $\alpha_0 = 0.05$. The second complication with power is that the expected α at a given level α' is no longer equal to α', and is generally unknown. Therefore we replace the mean and RMS of the error $\alpha - \alpha'$ from the size summaries with the simple mean and standard deviation of α over the same range.

We also evaluate the correlations among different permutation testing methods (most notably between the hypothetical exact test and the other tests). Firstly in terms of their sets of p-values from the N_s simulations, and secondly, averaged over multiple simulations, in terms of their sets of statistic values from the N_p permutations.

Results and discussion

The following series of tables provides detailed results for the 11 permutation methods summarised in table 2.1 together with results from the parametric F-statistic (after transformation in the case of $m > 1$, as described in section A.4.4). The third column heading indicates the distribution of the simulated error $\mathcal{E}$, with z denoting standard normal, t_5 a t-distribution on five degrees of freedom, and e^3 a cubed standard exponential distribution. The fourth column indicates the type of design X, as detailed in table 2.2.

Table 2.3 shows the observed false-positive rate at the 5% level. A 95% confidence interval based on the 10,000 simulations performed is (4.57, 5.43) (see table 2.13 later). Values outside the confidence interval are starred, and the numbers of values which are above it, i.e. significantly anti-conservative, are summarised at the bottom of the table. It is immediately obvious that Kennedy's method is unsuitable as a permutation test, since it completely fails to control type-I error, as already reported in [8], and explained earlier and in [40]. More interesting, are the findings that ter Braak's method and Shuffle-Y are also quite frequently too liberal, in particular, they both appear quite unstable with small n and highly-skewed errors; a discrete interest covariate leads to particularly bad results in some but not all cases. For $n \geq 9$ these methods appear valid, with tB appearing conservative for $n = 16$ and exp^3 errors. Note that $n = 9$ would typically still be considered a very low degree-of-freedom example (with $DF_E = 7$); the generally good performance of the methods suggests that Freedman and Lane (as quoted in [8], p.278) may have been unnecessarily wary in suggesting that the sample size 'should be relatively large'. Note that the zeros present for some methods with $n = 6$ and a discrete interest covariate are a result of the very low number of permutations available in these

[35] Some authors reserve this symbol for a rejection rate under H_0, where it is the false positive rate or type-I error, but it is essentially the same measure for a permutation test: the proportion of p-values less than or equal to some level, α'.

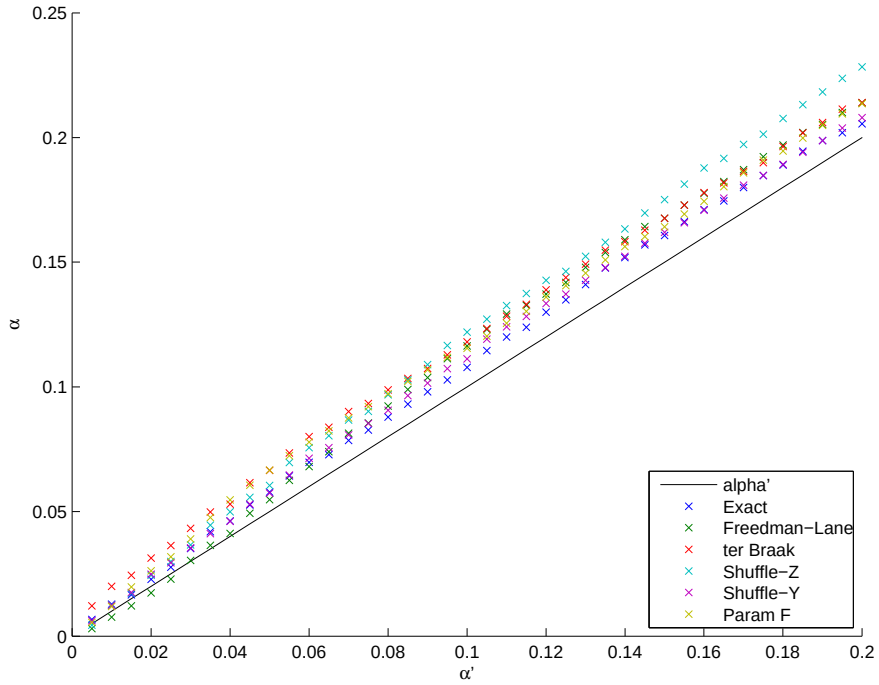
n	m	$\mathcal{E}$	X	Ex	FL	tB	HJ	Th	Re	SZ	Sm	SY	AY	Ke	PF
6	1	z	CC	4.86	4.97	5.01	3.92*	4.12*	4.1*	4.63	4.56*	5.76*	5.19	9.47*	4.79
6	1	z	DC	4.9	4.88	5.77*	4.17*	4.23*	3.69*	0*	4.24*	6.58*	0*	10.67*	4.96
6	1	z	CD	4.61	4.58	4.63	4.17*	3.75*	4.22*	4.58	4.49*	4.63	4.65	8.87*	4.36*
6	1	z	DD	4.72	5.02	5.22	3.72*	0*	2.11*	0*	2.39*	4.59	0*	7.22*	5.09
6	1	t_5	CC	5.3	5.27	5.54*	4.33*	4.28*	4.54*	5.09	5	6.03*	5.63*	10.38*	5.17
6	1	t_5	DC	4.98	4.99	6.27*	4.18*	4.34*	3.76*	0*	4.26*	6.55*	0*	10.58*	5.29
6	1	t_5	CD	5	4.84	4.96	4.09*	4.32*	4.02*	4.88	4.94	5.05	5.09	9.23*	4.89
6	1	t_5	DD	4.22*	4.61	5.17	3.47*	0*	1.99*	0*	2.26*	4.42*	0*	6.8*	4.89
6	1	e^3	CC	5.9*	5.63*	6.78*	5.31	5.2	4.98	6.29*	5.82*	5.91*	6.48*	11.61*	6.66*
6	1	e^3	DC	5.29	7.62*	11.13*	6.15*	7.11*	5.41	0.03*	6.68*	7.35*	0*	12.44*	9.66*
6	1	e^3	CD	5.56*	4.23*	5.25	4.5*	4.58	4.77	5.59*	5.83*	4.83	5.82*	9.97*	5.05
6	1	e^3	DD	5.07	5.82*	7.46*	4.66	0.18*	3.77*	0.08*	4.28*	5.48*	0*	7.02*	7.19*
9	1	z	CC	5.21	5.18	5.26	5.21	5.25	5.08	5.24	5.28	5.35	5.11	13.93*	5.21
9	1	z	DC	5.12	5.15	5.14	4.98	4.94	5.22	4.82	5.01	5.35	5.23	13.55*	4.93
9	1	z	CM	5.05	5.02	4.98	4.96	4.97	4.76	4.87	5	5.4	4.57*	13.36*	4.94
9	1	z	DM	5.24	5.31	5.47*	5.41	5.29	5.31	5.06	5.26	5.75*	4.9	14.23*	5.3
9	1	z	CD	5.19	5.11	5.1	5.05	5.05	5.03	5.23	5.1	5.12	5.13	13.44*	5.13
9	1	z	DD	4.99	4.75	4.82	4.67	4.77	4.82	4.93	4.79	4.83	5.04	13.71*	4.84
9	2	z	CC	4.87	4.97	4.88	4.83	4.97	5.04	4.86	4.94	5.06	5.34	19.81*	4.82
9	2	z	DC	4.8	5	5.09	4.74	4.74	4.84	4.74	4.85	4.87	5.16	19.67*	4.89
9	2	z	CM	4.91	4.98	4.97	4.96	4.97	5.14	4.93	5	5	4.75	20.11*	4.89
9	2	z	DM	5.35	5.66*	5.57*	5.33	5.32	5.19	5.17	5.24	5.48*	4.84	21.21*	5.44*
9	2	z	CD	5	5.06	5	4.92	5.36	5.18	4.94	4.95	5.02	4.91	19.86*	4.97
9	2	z	DD	4.97	4.8	4.74	4.97	4.88	4.93	5.02	4.65	4.81	5.18	19.98*	4.77
16	1	e^3	CC	4.94	4.76	4.39*	4.2*	4.89	3.99*	5.71*	4.37*	4.86	4.79	7.72*	4.27*
16	1	e^3	DC	4.97	4.9	4.46*	4.32*	5.01	3.86*	7.16*	4.59	4.93	4.76	8.16*	4.21*
16	1	e^3	CM	4.96	4.93	4.5*	4.21*	5.17	4.04*	5.43*	5.14	4.91	5.01	8.63*	4.33*
16	1	e^3	DM	5.22	5.17	4.8	4.6	5.56*	3.56*	6.31*	5.73*	5.07	5.43*	9.09*	4.53*
16	1	e^3	CD	4.65	4.05*	3.71*	3.6*	4.15*	4.22*	4.74	4.26*	4.63	4.67	7.75*	3.56*
16	1	e^3	DD	4.83	4.13*	3.41*	3.43*	3.95*	4.02*	5.12	4.19*	4.78	5.15	7.49*	3.27*
				2	4	8	1	2	0	6	4	9	4	30	4

Table 2.3: Accuracy, quantified by $100\,\alpha : \alpha' = 5\%$. Values outside the theoretical 95% confidence interval are starred. The final row is a count of the number of times α exceeded the upper confidence limit.

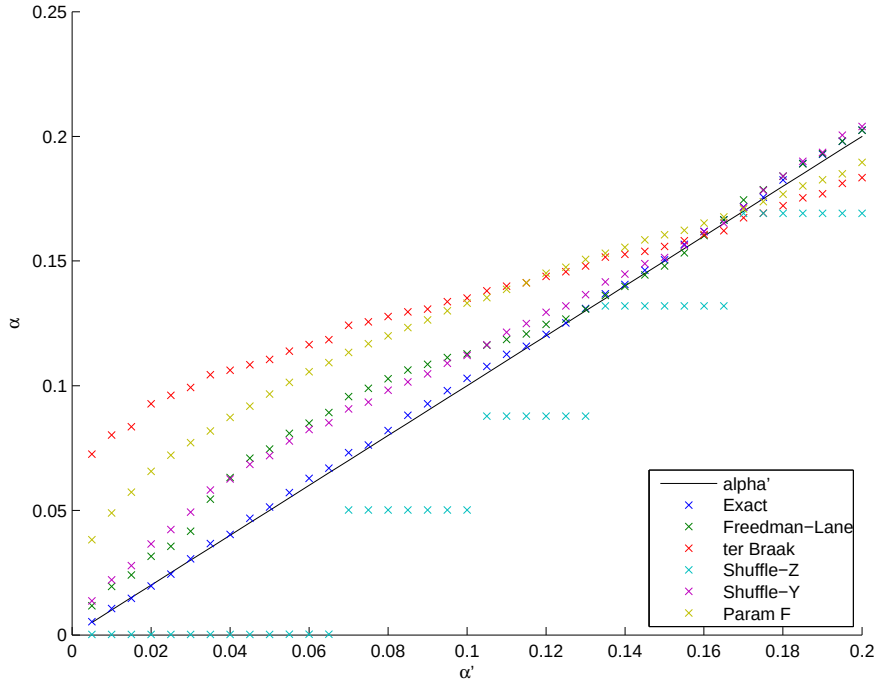
cases. Since the covariate is binary, there are at most ${}_6C_3 = 20$ distinct permutations of the interest available, which would mean that a p-value of 0.05 is the lowest achievable. This situation is particularly extreme, and arguably not of great practical importance. Nevertheless, it is of interest here, within the context of comparing different methods, because not all methods suffer to the same extent. In particular, Smith's method, which involves a relatively minor modification to Shuffle-Z, achieves results much closer to the expected values.

A closer look at the rejection rates is given in figure 2.1. As mentioned earlier, the validity of the method can depend on the level chosen, for example figure 2.1(b) shows tB initially severely anti-conservative, but becoming less liberal for higher α', eventually becoming conservative over $\alpha' = 0.17$. Table 2.4 therefore averages the error in test size over a range of the most important p-values. The general patterns are similar to table 2.3, including the conclusion that Ke is invalid, and that tB and SY can be somewhat erratic. However, the final summary row paints a slightly different picture because it considers departures in either direction from the expected rejection rate to be equally bad in terms of the ranking, unlike table 2.3 which ignores significantly conservative results. For this reason, table 2.4 shows the transformed residual strategies in a bad light — they have the worst rankings after Kennedy's method. However, as noted earlier in section 2.5 conservatism under the null need not imply low power for the alternative hypothesis. The four significantly liberal results for FL, compared to 0–2 for the transformed-residual strategies gives empirical support to Huh and Jhun's theoretical argument that their test is exact, while FL is only approximate due to its use of inexchangeable errors [50]. Anderson and Legendre observed from their simulations that 'when Manly's method [SY] gave too many rejections, ter Braak's method and the normal-theory t-test gave too few and vice versa' [8]. However, neither table 2.3 nor 2.4 seem to exhibit this trend; for example table comparing the signs of $\alpha - \alpha'$ for tB and SY in the two tables shows that they agree in 22 and 23 of the 30 cases, including the cases where they are (both) least accurate.

Although the large number of values present in the tables makes them somewhat tedious to interpret, it is important to note that simply averaging over the rows of the table loses potentially important information, since the relative performance of the different algorithms varies for different error distributions, types of covariate, or degrees of freedom. This means that the arbitrary proportions of e.g. normal or exponential-cubed errors considered for simulation will affect the summaries at the bottom of each table, which are intended only to give a first-impression of the results. In the interests of further simplifying the interpretation, but with the aforementioned caveat, figure 2.2 presents boxplots summarising the columns of table 2.4. The most interesting aspect of the boxplot is that it highlights the outliers present for some of the methods. The summary rows in the two tables made the parametric normal-theory test seem surprisingly accurate, given the high proportion of non-normal error distributions included: it matched the best of the realisable permutation methods (FL) in that both had four significantly liberal results, and they had almost equal average ranks. However, figure 2.2 shows that when PF breaks down, it can do so quite dramatically, whereas FL has only a single relatively



(a) Rejection rate, for design CC



(b) Rejection rate, for design DC

Figure 2.1: The observed false positive rate α plotted against the expected value under the null hypothesis α', for a subset of the methods and two different designs, with $n = 6$ and cubed-exponential errors.

n	m	$\mathcal{E}$	X	Ex	FL	tB	HJ	Th	Re	SZ	Sm	SY	AY	Ke	PF
6	1	z	CC	-0.21^{5}	-0.17^{3}	0.09^{2}	-1.9^{11}	-1.8^{9}	-1.8^{10}	-0.24^{6}	-0.35^{7}	0.57^{8}	0.033^{1}	2.7^{12}	-0.19^{4}
6	1	z	DC	-0.15^{2}	-0.58^{3}	1.6^{6}	-1.8^{8}	-1.8^{7}	-1.9^{9}	-2.5^{12}	-1^{4}	1^{5}	-2.5^{12}	2.4^{10}	-0.056^{1}
6	1	z	CD	-0.33^{6}	-0.36^{8}	-0.13^{1}	-1.8^{10}	-1.9^{11}	-1.8^{9}	-0.34^{7}	-0.32^{5}	-0.25^{3}	-0.24^{2}	2.2^{12}	-0.31^{4}
6	1	z	DD	-0.74^{4}	-0.62^{3}	1.3^{6}	-1.9^{7}	-2.5^{11}	-2.2^{9}	-2.5^{11}	-2^{8}	-0.76^{5}	-2.5^{11}	0.55^{2}	-0.0034^{1}
6	1	t_5	CC	-0.019^{2}	-0.097^{3}	0.46^{7}	-1.8^{10}	-1.8^{11}	-1.7^{9}	-0.14^{4}	-0.15^{5}	0.69^{8}	0.25^{6}	3.3^{12}	0.017^{1}
6	1	t_5	DC	-0.027^{1}	-0.48^{3}	1.9^{9}	-1.8^{7}	-1.8^{6}	-1.9^{8}	-2.5^{10}	-0.95^{4}	1.1^{5}	-2.5^{10}	2.9^{12}	0.17^{2}
6	1	t_5	CD	-0.2^{8}	-0.19^{6}	0.06^{1}	-1.8^{10}	-1.8^{9}	-1.8^{11}	-0.19^{7}	-0.15^{4}	-0.15^{5}	-0.089^{2}	2.6^{12}	-0.12^{3}
6	1	t_5	DD	-0.88^{5}	-0.82^{3}	1.3^{6}	-1.9^{7}	-2.5^{11}	-2.2^{9}	-2.5^{11}	-2^{8}	-0.83^{4}	-2.5^{11}	0.49^{2}	-0.028^{1}
6	1	e^3	CC	0.41^{3}	-0.0058^{1}	1.2^{8}	-1.6^{9}	-1.6^{10}	-1.7^{11}	0.56^{5}	0.21^{2}	0.44^{4}	0.67^{6}	3.9^{12}	0.79^{7}
6	1	e^3	DC	0.05^{1}	1.4^{4}	6.6^{12}	-1.4^{5}	-1.3^{3}	-1.6^{6}	-2.5^{8}	0.65^{2}	1.7^{7}	-2.5^{9}	4.7^{11}	4.3^{10}
6	1	e^3	CD	0.42^{6}	-0.44^{7}	0.26^{3}	-1.8^{11}	-1.7^{10}	-1.7^{9}	0.38^{5}	0.55^{8}	-0.04^{1}	0.32^{4}	2.8^{12}	0.11^{2}
6	1	e^3	DD	-0.47^{3}	0.61^{4}	4.2^{12}	-1.7^{6}	-2.4^{8}	-1.8^{7}	-2.4^{9}	-0.37^{2}	0.16^{1}	-2.5^{10}	1.4^{5}	3.1^{11}
9	1	z	CC	0.057^{2}	0.083^{6}	0.14^{10}	0.069^{3}	0.017^{1}	0.071^{4}	0.11^{9}	0.085^{7}	0.32^{11}	0.09^{8}	5.8^{12}	0.078^{5}
9	1	z	DC	0.031^{3}	0.077^{7}	0.21^{10}	-0.12^{8}	-0.027^{2}	-0.054^{6}	-0.032^{4}	0.033^{5}	0.3^{11}	0.17^{9}	5.7^{12}	-0.0052^{1}
9	1	z	CM	0.032^{6}	0.0039^{1}	0.072^{7}	-0.1^{9}	-0.024^{5}	-0.096^{8}	0.022^{4}	-0.012^{2}	0.35^{11}	-0.27^{10}	5.6^{12}	0.021^{3}
9	1	z	DM	0.24^{9}	0.24^{8}	0.44^{10}	0.17^{6}	0.054^{1}	0.088^{4}	0.074^{3}	0.14^{5}	0.69^{11}	-0.058^{2}	6.2^{12}	0.2^{7}
9	1	z	CD	0.087^{7}	0.083^{6}	0.073^{4}	-0.042^{2}	-0.0094^{1}	-0.078^{5}	0.12^{10}	0.094^{9}	0.12^{11}	0.049^{3}	5.5^{12}	0.092^{8}
9	1	z	DD	-0.1^{5}	-0.13^{7}	-0.054^{2}	-0.28^{9}	-0.37^{11}	-0.31^{10}	-0.12^{6}	-0.19^{8}	-0.091^{4}	0.027^{1}	5.6^{12}	-0.091^{3}
9	2	z	CC	-0.027^{6}	0.027^{4}	0.02^{3}	-0.16^{10}	-0.047^{7}	-0.09^{8}	0.002^{1}	-0.018^{2}	0.1^{9}	0.25^{11}	9.9^{12}	-0.027^{5}
9	2	z	DC	-0.086^{6}	-0.0081^{1}	0.074^{5}	-0.09^{7}	-0.2^{10}	-0.14^{8}	-0.16^{9}	-0.06^{4}	-0.054^{3}	0.22^{11}	10^{12}	-0.044^{2}
9	2	z	CM	-0.017^{7}	0.018^{8}	-0.0096^{3}	-0.11^{10}	-0.14^{11}	-0.013^{4}	-0.0035^{2}	0.00072^{1}	0.03^{9}	-0.013^{5}	10^{12}	-0.016^{6}
9	2	z	DM	0.13^{7}	0.31^{11}	0.27^{10}	0.099^{5}	0.12^{6}	0.014^{1}	0.033^{3}	0.098^{4}	0.17^{9}	0.031^{2}	11^{12}	0.15^{8}
9	2	z	CD	0.022^{4}	0.031^{6}	-0.029^{5}	-0.18^{11}	0.069^{9}	-0.095^{10}	0.0073^{2}	0.037^{7}	0.00032^{1}	-0.042^{8}	10^{12}	0.013^{3}
9	2	z	DD	-0.15^{3}	-0.15^{5}	-0.21^{9}	-0.11^{2}	-0.22^{10}	-0.18^{8}	-0.15^{4}	-0.26^{11}	-0.16^{6}	-0.041^{1}	10^{12}	-0.16^{7}
16	1	e^3	CC	-0.058^{3}	-0.22^{5}	-0.38^{6}	-0.41^{7}	-0.019^{1}	-0.53^{10}	0.54^{11}	-0.44^{8}	-0.11^{4}	-0.053^{2}	1.6^{12}	-0.45^{9}
16	1	e^3	DC	-0.039^{2}	-0.075^{3}	-0.3^{7}	-0.41^{8}	0.092^{4}	-0.69^{10}	1.6^{11}	-0.29^{6}	-0.12^{5}	-0.026^{1}	1.9^{12}	-0.45^{9}
16	1	e^3	CM	-0.076^{3}	-0.16^{6}	-0.38^{8}	-0.53^{10}	0.087^{4}	-0.62^{11}	0.23^{7}	-0.055^{2}	-0.15^{5}	-0.0082^{1}	2.1^{12}	-0.46^{9}
16	1	e^3	DM	0.092^{4}	0.37^{7}	0.28^{6}	0.078^{3}	0.66^{9}	-0.79^{10}	0.96^{11}	0.61^{8}	0.021^{1}	0.15^{5}	2.9^{12}	0.062^{2}
16	1	e^3	CD	-0.26^{2}	-0.64^{8}	-0.84^{9}	-0.9^{10}	-0.48^{5}	-0.54^{6}	-0.3^{3}	-0.56^{7}	-0.31^{4}	-0.15^{1}	1.4^{12}	-0.91^{11}
16	1	e^3	DD	-0.083^{2}	-0.63^{6}	-0.97^{9}	-1^{10}	-0.62^{5}	-0.69^{8}	0.029^{1}	-0.63^{7}	-0.17^{4}	0.12^{3}	1.4^{12}	-1.1^{11}
				4.233	5.1	6.533	7.7	6.933	7.933	6.533	5.4	5.833	5.6	11	5.2

Table 2.4: Average accuracy, quantified by $100\,\mathrm{mean}(\alpha - \alpha') : \alpha' \leq 5\%$, negative values are conservative. Superscripts show ranks (with 1 = best) based on distance from zero. The final row shows the mean rank for each method.

modest outlier. Interestingly, Smith's method performs very similarly to FL, and has no outlying liberal results. Once again, the validity (albeit with excessive conservatism) of the transformed residual strategies is evident.

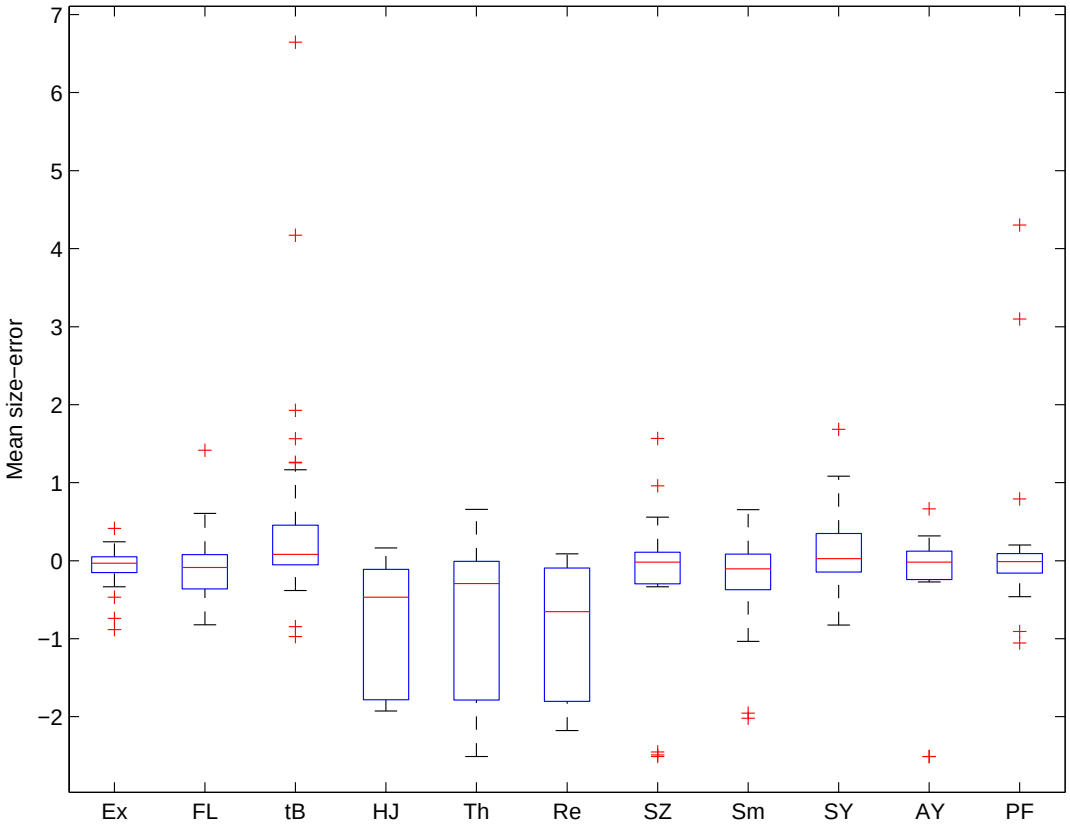


Figure 2.2: Boxplot showing the distribution of $100\,\text{mean}(\alpha - \alpha') : \alpha' \leq 5\%$ under the null hypothesis, over the 30 scenarios in the rows of table 2.4.

Tables 2.5 and 2.6 explore the variability in size around the expected level, and the uniformity of the p-values, respectively. These metrics are of secondary importance to accuracy, but are useful for uncovering and further understanding the differences between similarly accurate strategies. For example, FL and Sm are barely separable in terms of average accuracy, but FL appears to be slightly more successful in terms of having less variable accuracy and more uniform p-values. The transformed-residual strategies fair very poorly here. The most obvious explanation for this is that they use fewer permutations; even though the extra permutations used in the other methods are arguably exchanging inexchangeable errors, they could still lead to lower overall variability. To shed further light on this, figure 2.3 plots a histogram of the p-values for some of the methods in a particular situation. It seems that the three transformed-residual strategies suffer from a similar discreteness of the p-values due to the low number of reduced-space permutations, rather than favouring any particular range of p-values at the expense of the most relevant ones within the lowest bin.

Moving on to power, table 2.7 reports the rejection rates at the 5% level, and table 2.8 averages over a range of levels below this. Figure 2.4 summarises the distribution of average powers. Kennedy's method has been removed from consideration due to its invalid test

n	m	$\mathcal{E}$	X	Ex	FL	tB	HJ	Th	Re	SZ	Sm	SY	AY	Ke	PF
6	1	z	CC	0.228^{5}	0.183^{3}	0.164^{2}	2.22^{11}	2.21^{9}	2.21^{10}	0.267^{6}	0.363^{7}	0.604^{8}	0.0704^{1}	2.98^{12}	0.204^{4}
6	1	z	DC	0.188^{2}	0.622^{3}	1.63^{6}	2.21^{8}	2.21^{7}	2.23^{9}	$2.9^{10.5}$	1.1^{4}	1.14^{5}	$2.9^{10.5}$	3.03^{12}	0.0778^{1}
6	1	z	CD	0.358^{5}	0.395^{8}	0.261^{2}	2.21^{10}	2.23^{11}	2.21^{9}	0.371^{7}	0.367^{6}	0.27^{3}	0.257^{1}	2.4^{12}	0.356^{4}
6	1	z	DD	0.795^{3}	0.696^{2}	1.38^{6}	2.23^{8}	2.9^{11}	2.42^{9}	2.9^{11}	2.16^{7}	0.821^{4}	2.9^{11}	1.06^{5}	0.0809^{1}
6	1	t_5	CC	0.151^{2}	0.159^{3}	0.481^{7}	2.2^{10}	2.2^{11}	2.2^{9}	0.161^{4}	0.162^{5}	0.739^{8}	0.289^{6}	3.57^{12}	0.0874^{1}
6	1	t_5	DC	0.0685^{1}	0.508^{3}	1.97^{6}	2.21^{8}	2.2^{7}	2.23^{9}	$2.9^{10.5}$	1.02^{4}	1.17^{5}	$2.9^{10.5}$	3.42^{12}	0.188^{2}
6	1	t_5	CD	0.219^{8}	0.207^{6}	0.157^{3}	2.21^{10}	2.2^{9}	2.21^{11}	0.216^{7}	0.167^{5}	0.162^{4}	0.121^{1}	2.85^{12}	0.132^{2}
6	1	t_5	DD	0.943^{5}	0.882^{2}	1.36^{6}	2.25^{8}	2.9^{11}	2.45^{9}	2.9^{11}	2.24^{7}	0.883^{3}	2.9^{11}	0.936^{4}	0.0428^{1}
6	1	e^3	CC	0.484^{3}	0.25^{1}	1.22^{8}	2.22^{11}	2.22^{10}	2.21^{9}	0.683^{5}	0.365^{2}	0.496^{4}	0.782^{6}	4.3^{12}	0.964^{7}
6	1	e^3	DC	0.108^{1}	1.59^{3}	6.74^{12}	2.26^{6}	2.42^{7}	2.21^{5}	2.88^{8}	0.842^{2}	1.79^{4}	2.9^{9}	5.13^{11}	4.35^{10}
6	1	e^3	CD	0.433^{6}	0.525^{7}	0.276^{3}	2.2^{10}	2.2^{9}	2.2^{11}	0.408^{5}	0.593^{8}	0.144^{2}	0.372^{4}	3.11^{12}	0.126^{1}
6	1	e^3	DD	0.547^{1}	0.993^{4}	4.32^{12}	2.13^{6}	2.82^{8}	2.18^{7}	2.84^{9}	0.579^{3}	0.559^{2}	2.9^{10}	1.84^{5}	3.13^{11}
9	1	z	CC	0.0976^{1}	0.107^{2}	0.15^{10}	0.117^{5}	0.113^{4}	0.134^{8}	0.127^{7}	0.136^{9}	0.356^{11}	0.11^{3}	6.21^{12}	0.124^{6}
9	1	z	DC	0.0588^{2}	0.0934^{7}	0.228^{10}	0.136^{8}	0.071^{4}	0.0929^{6}	0.0777^{5}	0.059^{3}	0.312^{11}	0.194^{9}	6.11^{12}	0.0343^{1}
9	1	z	CM	0.0489^{2}	0.0343^{1}	0.0839^{6}	0.119^{8}	0.0897^{7}	0.126^{9}	0.0558^{5}	0.052^{3}	0.369^{11}	0.318^{10}	5.96^{12}	0.0531^{4}
9	1	z	DM	0.271^{9}	0.266^{8}	0.456^{10}	0.208^{6}	0.1^{3}	0.12^{4}	0.094^{2}	0.176^{5}	0.717^{11}	0.0854^{1}	6.54^{12}	0.23^{7}
9	1	z	CD	0.109^{8}	0.106^{7}	0.0823^{3}	0.0725^{1}	0.0827^{4}	0.0921^{5}	0.142^{11}	0.118^{9}	0.139^{10}	0.0729^{2}	5.85^{12}	0.106^{6}
9	1	z	DD	0.11^{5}	0.141^{7}	0.0992^{2}	0.308^{9}	0.396^{11}	0.332^{10}	0.131^{6}	0.192^{8}	0.102^{4}	0.0686^{1}	6.05^{12}	0.1^{3}
9	2	z	CC	0.0647^{3}	0.0615^{2}	0.0849^{4}	0.196^{10}	0.0892^{6}	0.107^{8}	0.0855^{5}	0.0585^{1}	0.127^{9}	0.297^{11}	10.5^{12}	0.0972^{7}
9	2	z	DC	0.112^{6}	0.0385^{1}	0.083^{3}	0.134^{7}	0.228^{10}	0.15^{8}	0.188^{9}	0.0928^{5}	0.0921^{4}	0.239^{11}	10.6^{12}	0.0767^{2}
9	2	z	CM	0.0718^{7}	0.0391^{2}	0.0601^{4}	0.127^{9}	0.163^{11}	0.0672^{6}	0.0736^{8}	0.0373^{1}	0.0546^{3}	0.139^{10}	10.9^{12}	0.061^{5}
9	2	z	DM	0.171^{6}	0.373^{11}	0.309^{10}	0.177^{7}	0.158^{5}	0.0997^{3}	0.0619^{1}	0.149^{4}	0.222^{9}	0.0841^{2}	11.8^{12}	0.198^{8}
9	2	z	CD	0.0651^{7}	0.0616^{5}	0.0543^{2}	0.19^{11}	0.146^{9}	0.157^{10}	0.0561^{3}	0.0632^{6}	0.0679^{8}	0.0585^{4}	10.8^{12}	0.0463^{1}
9	2	z	DD	0.177^{5}	0.175^{4}	0.229^{9}	0.13^{2}	0.239^{10}	0.199^{8}	0.171^{3}	0.282^{11}	0.182^{6}	0.102^{1}	10.8^{12}	0.185^{7}
16	1	e^3	CC	0.0933^{2}	0.236^{5}	0.419^{6}	0.454^{7}	0.12^{3}	0.607^{11}	0.59^{10}	0.487^{8}	0.125^{4}	0.0886^{1}	1.81^{12}	0.5^{9}
16	1	e^3	DC	0.0594^{1}	0.0853^{2}	0.33^{7}	0.466^{8}	0.128^{3}	0.764^{10}	1.67^{11}	0.313^{6}	0.141^{5}	0.14^{4}	2.03^{12}	0.504^{9}
16	1	e^3	CM	0.0953^{2}	0.168^{6}	0.416^{8}	0.586^{10}	0.118^{3}	0.682^{11}	0.278^{7}	0.119^{4}	0.158^{5}	0.0581^{1}	2.33^{12}	0.507^{9}
16	1	e^3	DM	0.109^{2}	0.377^{7}	0.343^{6}	0.291^{5}	0.67^{9}	0.902^{10}	1^{11}	0.626^{8}	0.063^{1}	0.195^{3}	3.07^{12}	0.285^{4}
16	1	e^3	CD	0.276^{2}	0.681^{8}	0.921^{9}	0.987^{10}	0.54^{5}	0.592^{7}	0.311^{3}	0.591^{6}	0.327^{4}	0.174^{1}	1.62^{12}	1^{11}
16	1	e^3	DD	0.0983^{2}	0.67^{5}	1.08^{9}	1.11^{10}	0.677^{7}	0.753^{8}	0.0927^{1}	0.674^{6}	0.176^{4}	0.132^{3}	1.64^{12}	1.17^{11}
				3.8	4.5	6.367	7.967	7.467	8.3	6.733	5.433	5.733	5.3	11.23	5.167

Table 2.5: Size variability, quantified by $100\sqrt{\text{mean}((\alpha-\alpha')^2)} : \alpha' \leq 5\%$, smaller is better. Superscripts show ranks, with 1 = best. The final row shows the mean rank for each method.

n	m	$\mathcal{E}$	X	Ex	FL	tB	HJ	Th	Re	SZ	Sm	SY	AY	Ke	PF
6	1	z	CC	0.059^{4}	0.064^{5}	0.09^{7}	0.83^{11}	0.83^{11}	0.83^{11}	0.047^{2}	0.07^{6}	0.095^{8}	0.044^{1}	0.11^{9}	0.056^{3}
6	1	z	DC	0.048^{2}	0.16^{5}	0.37^{7}	0.83^{9}	0.83^{9}	0.83^{9}	Inf^{12}	0.27^{6}	0.11^{4}	Inf^{12}	0.1^{3}	0.035^{1}
6	1	z	CD	0.057^{6}	0.059^{7}	0.083^{9}	0.83^{11}	0.83^{11}	0.83^{11}	0.054^{5}	0.044^{2}	0.049^{4}	0.04^{1}	0.07^{8}	0.047^{3}
6	1	z	DD	0.22^{3}	0.27^{5}	0.47^{7}	$0.83^{8.5}$	Inf^{11}	$0.83^{8.5}$	Inf^{11}	0.43^{6}	0.21^{2}	Inf^{11}	0.26^{4}	0.031^{1}
6	1	t_5	CC	0.063^{5}	0.076^{6}	0.089^{7}	0.83^{11}	0.83^{11}	0.83^{11}	0.06^{4}	0.055^{3}	0.1^{8}	0.04^{2}	0.11^{9}	0.039^{1}
6	1	t_5	DC	0.044^{2}	0.14^{5}	0.37^{7}	0.83^{9}	0.83^{9}	0.83^{9}	Inf^{12}	0.22^{6}	0.099^{4}	Inf^{12}	0.075^{3}	0.027^{1}
6	1	t_5	CD	0.068^{7}	0.038^{2}	0.081^{8}	0.83^{11}	0.83^{11}	0.83^{11}	0.054^{4}	0.053^{3}	0.06^{5}	0.065^{6}	0.11^{9}	0.029^{1}
6	1	t_5	DD	0.19^{2}	0.27^{5}	0.46^{7}	$0.83^{8.5}$	Inf^{11}	$0.83^{8.5}$	Inf^{11}	0.44^{6}	0.2^{3}	Inf^{11}	0.25^{4}	0.023^{1}
6	1	e^3	CC	0.037^{2}	0.1^{7}	0.11^{8}	0.83^{11}	0.83^{11}	0.83^{11}	0.048^{5}	0.076^{6}	0.047^{4}	0.044^{3}	0.12^{9}	0.034^{1}
6	1	e^3	DC	0.043^{1}	0.074^{2}	0.6^{7}	0.83^{8}	0.83^{10}	0.83^{9}	1^{11}	0.11^{3}	0.12^{4}	Inf^{12}	0.14^{5}	0.31^{6}
6	1	e^3	CD	0.089^{8}	0.04^{1}	0.098^{9}	0.83^{11}	0.83^{11}	0.83^{11}	0.069^{4}	0.069^{5}	0.075^{6}	0.045^{2}	0.086^{7}	0.047^{3}
6	1	e^3	DD	0.19^{1}	0.27^{4}	0.78^{8}	0.81^{9}	0.61^{7}	0.82^{10}	1^{11}	0.28^{5}	0.25^{3}	Inf^{12}	0.24^{2}	0.45^{6}
9	1	z	CC	0.036^{6}	0.022^{2}	0.028^{4}	0.039^{8}	0.046^{9}	0.071^{11}	0.022^{1}	0.037^{7}	0.062^{10}	0.026^{3}	0.16^{12}	0.033^{5}
9	1	z	DC	0.03^{3}	0.024^{2}	0.064^{11}	0.045^{6}	0.051^{9}	0.051^{8}	0.051^{7}	0.031^{4}	0.058^{10}	0.037^{5}	0.15^{12}	0.023^{1}
9	1	z	CM	0.026^{2}	0.025^{1}	0.041^{6}	0.045^{8}	0.055^{10}	0.059^{11}	0.039^{5}	0.03^{3}	0.053^{9}	0.043^{7}	0.16^{12}	0.03^{4}
9	1	z	DM	0.057^{9}	0.036^{5}	0.069^{10}	0.039^{6}	0.034^{4}	0.042^{7}	0.024^{1}	0.025^{2}	0.089^{11}	0.026^{3}	0.17^{12}	0.044^{8}
9	1	z	CD	0.025^{4}	0.03^{5}	0.024^{3}	0.036^{10}	0.046^{11}	0.032^{7}	0.031^{6}	0.034^{8}	0.035^{9}	0.024^{2}	0.16^{12}	0.024^{1}
9	1	z	DD	0.041^{6}	0.037^{5}	0.034^{4}	0.054^{9}	0.081^{11}	0.061^{10}	0.049^{8}	0.047^{7}	0.026^{1}	0.029^{3}	0.14^{12}	0.028^{2}
9	2	z	CC	0.026^{1}	0.032^{2}	0.053^{10}	0.045^{8}	0.044^{7}	0.04^{5}	0.044^{6}	0.034^{3}	0.052^{9}	0.039^{4}	0.2^{12}	0.054^{11}
9	2	z	DC	0.025^{4}	0.02^{1}	0.026^{5}	0.069^{11}	0.047^{9}	0.037^{8}	0.023^{2}	0.024^{3}	0.03^{6}	0.054^{10}	0.21^{12}	0.031^{7}
9	2	z	CM	0.039^{8}	0.022^{1}	0.025^{2}	0.042^{9}	0.05^{10}	0.032^{6}	0.027^{4}	0.025^{3}	0.028^{5}	0.056^{11}	0.2^{12}	0.037^{7}
9	2	z	DM	0.027^{3}	0.023^{1}	0.038^{6}	0.054^{11}	0.039^{7}	0.042^{8}	0.025^{2}	0.034^{4}	0.035^{5}	0.051^{10}	0.22^{12}	0.042^{9}
9	2	z	CD	0.036^{6}	0.026^{1}	0.027^{2}	0.053^{9}	0.053^{10}	0.066^{11}	0.031^{5}	0.038^{8}	0.038^{7}	0.027^{3}	0.21^{12}	0.03^{4}
9	2	z	DD	0.06^{9}	0.037^{2}	0.044^{5}	0.034^{1}	0.063^{10}	0.037^{3}	0.07^{11}	0.056^{8}	0.05^{7}	0.048^{6}	0.21^{12}	0.039^{4}
16	1	e^3	CC	0.031^{2}	0.045^{4}	0.055^{8}	0.045^{5}	0.047^{6}	0.028^{1}	0.084^{12}	0.063^{11}	0.039^{3}	0.054^{7}	0.059^{10}	0.058^{9}
16	1	e^3	DC	0.028^{2}	0.024^{1}	0.042^{4}	0.048^{6}	0.051^{8}	0.055^{9}	0.11^{12}	0.04^{3}	0.051^{7}	0.055^{10}	0.065^{11}	0.042^{5}
16	1	e^3	CM	0.042^{4}	0.051^{6}	0.067^{8}	0.078^{12}	0.025^{1}	0.07^{9}	0.032^{2}	0.058^{7}	0.051^{5}	0.034^{3}	0.074^{10}	0.076^{11}
16	1	e^3	DM	0.022^{1}	0.087^{6}	0.12^{11}	0.11^{8}	0.11^{9}	0.057^{4}	0.091^{7}	0.075^{5}	0.026^{2}	0.031^{3}	0.14^{12}	0.11^{10}
16	1	e^3	CD	0.045^{4}	0.086^{9}	0.093^{10}	0.11^{12}	0.037^{2}	0.069^{7}	0.062^{6}	0.076^{8}	0.055^{5}	0.044^{3}	0.023^{1}	0.094^{11}
16	1	e^3	DD	0.019^{1}	0.086^{8}	0.11^{12}	0.11^{11}	0.066^{6}	0.08^{7}	0.034^{2}	0.093^{9}	0.034^{3}	0.035^{4}	0.046^{5}	0.1^{10}
				3.933	3.867	7.067	8.933	8.733	8.4	6.333	5.333	5.633	6.033	8.833	4.9

Table 2.6: P-value uniformity, quantified by KS statistic (p vs uniform) : $p \leq 5\%$, smaller values indicate more uniform p-values. Superscripts show ranks, with 1 = best. The final row shows the mean rank for each method.

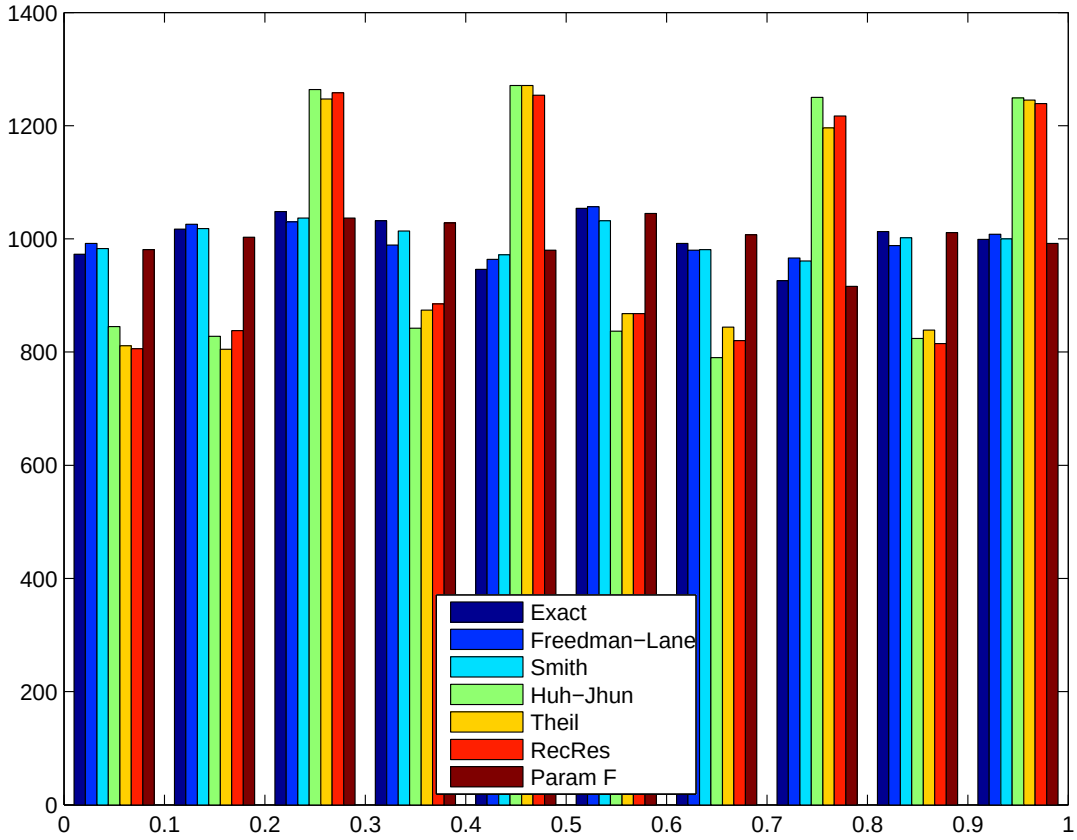


Figure 2.3: Histogram of the 10,000 simulated p-values under the null hypothesis, for $n = 6$, t-distributed errors, and continuous interest with discrete nuisance.

size; SY and tB have been retained, although their power results must be interpreted in the context of their occasional failures to control false positives. The most important question that arises from the tables is what it means for some of the realisable methods to have greater power than the hypothetical exact test. In particular, it is surprising that FL has a better average rank than the exact method that it approximates. Looking more closely for the sources of its superior rankings, we observe that FL beats Ex in terms of average power on 15 out of 30 occasions; of these, only 5 occur in situations for which FL has average size-error greater than that of Ex and greater than zero (table 2.4). It cannot be argued from these results that FL's apparently greater power arises from a failure to control size, though further investigation is clearly warranted to explain this surprising result. Performing the same exercise to compare ter Braak's method with FL leads to a more interesting conclusion. Of the 23 cases for which tB is more powerful, 16 show that it has a larger positive error in size compared to FL. Yet more damningly, the six scenarios for which tB has the greatest power advantages over FL and Ex (of over 10%) are precisely the same $n = 6$ designs with discrete interest for which it also has the largest average size error, and for which its average size is furthest from that of both FL and Ex. This suggests strongly that the method of ter Braak does not provide a legitimate power advantage over FL, but that rather it can achieve greater apparent power in certain cases due only to its loose control of false positives under challenging circumstances.

n	m	$\mathcal{E}$	X	Ex	FL	tB	HJ	Th	Re	SZ	Sm	SY	AY	PF
6	1	z	CC	44.13^{5}	44.59^{3}	45.28^{2}	25.21^{11}	25.93^{9}	25.8^{10}	43.81^{7}	44.01^{6}	44.31^{4}	40.07^{8}	45.66^{1}
6	1	z	DC	46.67^{6}	59.08^{4}	69.89^{2}	39.7^{7}	18.34^{9}	38.68^{8}	0.075^{10}	57.77^{5}	67.11^{3}	0^{11}	72.38^{1}
6	1	z	CD	43.7^{5}	43.42^{7}	43.91^{2}	24.86^{10}	25.06^{9}	24.75^{11}	43.84^{4}	43.51^{6}	43.85^{3}	40^{8}	45.3^{1}
6	1	z	DD	41.74^{5}	53.23^{3}	70.8^{2}	38.47^{6}	0.04^{10}	24.55^{8}	0.095^{9}	38.25^{7}	50.59^{4}	0^{11}	74.9^{1}
6	1	t_5	CC	37.41^{4}	37.79^{3}	38.71^{1}	21.31^{11}	21.79^{10}	21.84^{9}	36.95^{7}	37.02^{6}	37.2^{5}	35^{8}	38.64^{2}
6	1	t_5	DC	43.21^{6}	53.53^{4}	62.96^{2}	33.68^{8}	16.25^{9}	34.42^{7}	0.07^{10}	51.42^{5}	59.34^{3}	0^{11}	64.66^{1}
6	1	t_5	CD	36.29^{6}	36.14^{7}	36.65^{3}	21.31^{9}	21.29^{10}	20.76^{11}	36.41^{5}	36.44^{4}	36.75^{2}	34.29^{8}	38.02^{1}
6	1	t_5	DD	38.27^{5}	47.92^{3}	63.15^{2}	34.03^{6}	0.0325^{10}	21.24^{8}	0.055^{9}	33.53^{7}	45.27^{4}	0^{11}	66.76^{1}
6	1	e^3	CC	20.7^{3}	20.8^{2}	21.46^{1}	13.81^{11}	17.09^{9}	13.92^{10}	20.63^{4}	19.86^{7}	20.05^{6}	19.58^{8}	20.57^{5}
6	1	e^3	DC	21.15^{6}	26.66^{4}	33.92^{1}	19.87^{7}	15.11^{9}	19.29^{8}	0.305^{10}	25.76^{5}	28.99^{3}	0^{11}	30.81^{2}
6	1	e^3	CD	20.99^{2}	19.3^{7}	20^{6}	13.55^{11}	15.56^{9}	13.92^{10}	20.39^{4}	20.56^{3}	21.01^{1}	18.69^{8}	20.33^{5}
6	1	e^3	DD	20.72^{5}	24.09^{4}	31.27^{2}	18.83^{6}	0.735^{9}	13.03^{8}	0.405^{10}	17.55^{7}	25.55^{3}	0^{11}	31.29^{1}
9	1	z	CC	55.05^{7}	55.09^{6}	55.19^{3}	52.11^{8}	51.78^{9}	51.02^{10}	55.14^{5}	55.22^{2}	55.35^{1}	3.66^{11}	55.18^{4}
9	1	z	DC	69.32^{6}	69.8^{1}	69.76^{3}	65.41^{9}	66.9^{8}	64.44^{10}	68.03^{7}	69.55^{4}	69.39^{5}	3.333^{11}	69.77^{2}
9	1	z	CM	63.96^{4}	63.8^{6}	63.97^{3}	59.38^{10}	59.64^{9}	60.08^{8}	63.83^{5}	63.69^{7}	64^{2}	3.117^{11}	64^{1}
9	1	z	DM	82.97^{6}	83.4^{3}	83.48^{2}	81.08^{8}	80.32^{9}	79.34^{10}	81.11^{7}	83.24^{4}	83.14^{5}	10.35^{11}	83.63^{1}
9	1	z	CD	64.56^{7}	64.61^{6}	64.62^{5}	60.85^{8}	59.83^{10}	59.88^{9}	64.65^{3}	64.62^{4}	64.75^{2}	3.853^{11}	64.81^{1}
9	1	z	DD	84.61^{6}	85.05^{4}	85.06^{3}	83.14^{8}	80.92^{9}	80.23^{10}	83.63^{7}	84.7^{5}	85.15^{2}	6.825^{11}	85.47^{1}
9	2	z	CC	51.19^{5}	51.3^{3}	51.21^{4}	48.76^{8}	48.5^{9}	47.87^{10}	51.1^{7}	51.16^{6}	51.48^{1}	4.45^{11}	51.35^{2}
9	2	z	DC	64.67^{6}	66.11^{1}	66.09^{2}	61.86^{9}	62.94^{8}	61.06^{10}	63.65^{7}	65.44^{4}	65.11^{5}	4.317^{11}	65.78^{3}
9	2	z	CM	59.95^{3}	60.03^{1}	59.93^{4}	56.46^{9}	56.46^{10}	56.67^{8}	$59.92^{5.5}$	59.75^{7}	$59.92^{5.5}$	4.348^{11}	59.99^{2}
9	2	z	DM	79.37^{6}	81^{1}	80.35^{3}	78.27^{7}	77.29^{9}	76.91^{10}	77.93^{8}	80.25^{4}	80.15^{5}	6.272^{11}	80.92^{2}
9	2	z	CD	60.49^{2}	60.32^{5}	60.11^{7}	56.9^{8}	56.3^{10}	56.4^{9}	60.39^{4}	60.31^{6}	60.4^{3}	4.7^{11}	60.51^{1}
9	2	z	DD	80.82^{6}	81.78^{3}	81.17^{5}	79.66^{8}	78.06^{9}	77.64^{10}	80.35^{7}	81.54^{4}	81.8^{2}	5.218^{11}	82.28^{1}
16	1	e^3	CC	17.17^{4}	17.18^{3}	16.97^{6}	16.35^{9}	17.4^{2}	16.2^{10}	18.23^{1}	16.91^{7}	17.12^{5}	5.17^{11}	16.4^{8}
16	1	e^3	DC	$22.88^{4.5}$	23.13^{3}	$22.88^{4.5}$	21.73^{9}	23.19^{2}	21.17^{10}	25.7^{1}	22.77^{6}	22.65^{7}	5.713^{11}	21.83^{8}
16	1	e^3	CM	21.34^{4}	21.28^{5}	21.02^{7}	19.99^{9}	21.43^{3}	19.82^{10}	21.98^{1}	21.54^{2}	21.24^{6}	5.522^{11}	20.23^{8}
16	1	e^3	DM	33.58^{6}	33.79^{4}	33.79^{5}	32.1^{9}	34.43^{3}	30.56^{10}	35.61^{1}	34.65^{2}	33.42^{7}	13.09^{11}	32.19^{8}
16	1	e^3	CD	21.71^{1}	21.08^{5}	20.66^{8}	19.94^{10}	20.89^{6}	20.76^{7}	21.48^{3}	21.22^{4}	21.68^{2}	5.29^{11}	20.05^{9}
16	1	e^3	DD	35.28^{1}	34.25^{4}	34.05^{7}	32.45^{10}	34.08^{6}	33.62^{8}	35.22^{3}	34.23^{5}	35.25^{2}	13.1^{11}	32.66^{9}
				4.75	3.833	3.583	8.633	8.1	9.233	5.717	5.033	3.617	10.4	3.1

Table 2.7: Power, quantified by $100\,\alpha : \alpha' = 5\%$, higher is better. Superscripts show ranks, with 1 = best. The final row shows the mean rank for each method.

n	m	$\mathcal{E}$	X	Ex	FL	tB	HJ	Th	Re	SZ	Sm	SY	AY	PF
6	1	z	CC	27.26^{4}	27.07^{5}	30.8^{1}	4.235^{11}	4.356^{9}	4.334^{10}	26.77^{6}	26.32^{7}	27.8^{3}	25.05^{8}	29.96^{2}
6	1	z	DC	23.12^{6}	34.16^{4}	60.47^{1}	6.67^{7}	3.081^{9}	6.499^{8}	0.0003^{10}	28.59^{5}	45.51^{3}	0^{11}	56^{2}
6	1	z	CD	26.72^{6}	26.69^{7}	29.25^{2}	4.176^{10}	4.21^{9}	4.158^{11}	26.78^{5}	26.81^{4}	26.98^{3}	24.81^{8}	29.65^{1}
6	1	z	DD	16.77^{5}	29.91^{3}	61.63^{1}	6.462^{7}	0.00672^{9}	4.125^{8}	0.00038^{10}	15.16^{6}	23.14^{4}	0^{11}	58.37^{2}
6	1	t_5	CC	22.35^{4}	22.13^{5}	25.85^{1}	3.58^{11}	3.66^{10}	3.67^{9}	21.96^{6}	21.48^{7}	22.81^{3}	21.35^{8}	24.74^{2}
6	1	t_5	DC	21.47^{6}	29.57^{4}	53.7^{1}	5.658^{8}	2.731^{9}	5.783^{7}	0.00028^{10}	24.42^{5}	39.24^{3}	0^{11}	48.14^{2}
6	1	t_5	CD	21.42^{7}	21.46^{6}	23.83^{2}	3.58^{9}	3.577^{10}	3.487^{11}	21.5^{5}	21.7^{4}	21.93^{3}	20.55^{8}	24.08^{1}
6	1	t_5	DD	15.31^{5}	26.15^{3}	54.04^{1}	5.718^{7}	0.00546^{9}	3.569^{8}	0.00022^{10}	12.54^{6}	20.24^{4}	0^{11}	49.87^{2}
6	1	e^3	CC	13.98^{3}	13.94^{4}	15.53^{1}	2.329^{11}	2.88^{9}	2.347^{10}	13.93^{5}	13.03^{7}	13.84^{6}	12.54^{8}	14.58^{2}
6	1	e^3	DC	10.46^{6}	16.91^{4}	29.01^{1}	3.407^{7}	2.59^{9}	3.304^{8}	0.04853^{10}	15.07^{5}	20.04^{3}	0^{11}	23.98^{2}
6	1	e^3	CD	14.19^{3}	12.86^{7}	14.29^{2}	2.283^{11}	2.62^{9}	2.343^{10}	13.76^{6}	13.76^{5}	14.18^{4}	12.02^{8}	14.36^{1}
6	1	e^3	DD	8.63^{5}	13.64^{3}	27.3^{1}	3.265^{7}	0.1838^{9}	2.274^{8}	0.06636^{10}	8.038^{6}	13.26^{4}	0^{11}	24.39^{2}
9	1	z	CC	39.83^{5}	39.53^{7}	40.11^{3}	35.16^{8}	34.54^{9}	33.72^{10}	39.74^{6}	39.84^{4}	40.15^{1}	1.752^{11}	40.11^{2}
9	1	z	DC	53.31^{6}	55.03^{3}	55.74^{1}	46.83^{9}	48.86^{8}	45.53^{10}	51.85^{7}	54.52^{4}	53.97^{5}	1.442^{11}	55.32^{2}
9	1	z	CM	48.99^{4}	48.48^{7}	49.19^{1}	41.81^{9}	41.81^{8}	41.76^{10}	48.91^{5}	48.51^{6}	49.01^{3}	1.466^{11}	49.17^{2}
9	1	z	DM	68.68^{6}	71.31^{3}	73.04^{1}	65.14^{8}	63.25^{9}	62^{10}	66.44^{7}	71.08^{4}	70.61^{5}	3.926^{11}	72.85^{2}
9	1	z	CD	49.3^{5}	49.12^{7}	49.41^{3}	42.56^{8}	41.62^{9}	41.51^{10}	49.34^{4}	49.14^{6}	49.47^{2}	1.903^{11}	49.58^{1}
9	1	z	DD	70.16^{6}	72.69^{3}	74.17^{2}	66.78^{8}	61.06^{10}	63.26^{9}	69.41^{7}	71.99^{5}	72.29^{4}	1.866^{11}	74.62^{1}
9	2	z	CC	34.31^{4}	34.18^{6}	34.41^{3}	31.15^{8}	30.81^{9}	30.34^{10}	34.21^{5}	34.17^{7}	34.5^{1}	2.154^{11}	34.42^{2}
9	2	z	DC	45.89^{6}	48.6^{2}	48.72^{1}	42.29^{9}	43.9^{8}	41.52^{10}	45.26^{7}	47.67^{4}	47.11^{5}	2.175^{11}	48.24^{3}
9	2	z	CM	42.52^{4}	42.46^{6}	42.69^{1}	37.93^{8}	37.69^{9}	37.68^{10}	42.46^{5}	42.18^{7}	42.54^{3}	2.124^{11}	42.63^{2}
9	2	z	DM	61.68^{6}	65.57^{2}	65.22^{3}	60.27^{8}	58.68^{9}	57.9^{10}	60.54^{7}	64.77^{4}	64.38^{5}	3.316^{11}	66.2^{1}
9	2	z	CD	42.65^{4}	42.54^{5}	42.52^{6}	38.07^{8}	37.55^{9}	37.27^{10}	42.66^{3}	42.49^{7}	42.74^{2}	2.326^{11}	42.79^{1}
9	2	z	DD	63.13^{7}	66.16^{2}	65.3^{5}	61.25^{8}	56.66^{10}	58.85^{9}	63.18^{6}	65.69^{4}	65.89^{3}	2.464^{11}	67.83^{1}
16	1	e^3	CC	12.17^{3}	12.15^{4}	11.95^{6}	11.39^{9}	12.42^{2}	11.28^{10}	12.94^{1}	11.85^{7}	12.02^{5}	2.568^{11}	11.45^{8}
16	1	e^3	DC	16.87^{5}	17.18^{3}	17.12^{4}	15.95^{9}	17.27^{2}	15.53^{10}	19.16^{1}	16.85^{6}	16.66^{7}	2.935^{11}	16.11^{8}
16	1	e^3	CM	15.48^{3}	15.36^{5}	15.21^{7}	14.31^{9}	15.66^{2}	14.29^{10}	15.9^{1}	15.48^{4}	15.33^{6}	2.793^{11}	14.48^{8}
16	1	e^3	DM	26.63^{6}	26.76^{5}	27.08^{4}	25.35^{9}	27.36^{3}	23.85^{10}	28.2^{1}	27.48^{2}	26.47^{7}	6.947^{11}	25.46^{8}
16	1	e^3	CD	15.9^{1}	15.2^{6}	15.06^{7}	14.27^{10}	15.21^{5}	15.03^{8}	15.57^{3}	15.26^{4}	15.84^{2}	2.629^{11}	14.43^{9}
16	1	e^3	DD	28.18^{1}	27.14^{5}	27.38^{4}	25.53^{10}	27.08^{7}	26.71^{8}	28.01^{3}	27.11^{6}	28.12^{2}	6.472^{11}	25.91^{9}
				4.733	4.533	2.567	8.7	7.933	9.4	5.733	5.267	3.7	10.4	3.033

Table 2.8: Average power, quantified by $100\,\text{mean}(\alpha) : \alpha' \leq 5\%$, higher is better. Superscripts show ranks, with 1 = best. The final row shows the mean rank for each method.

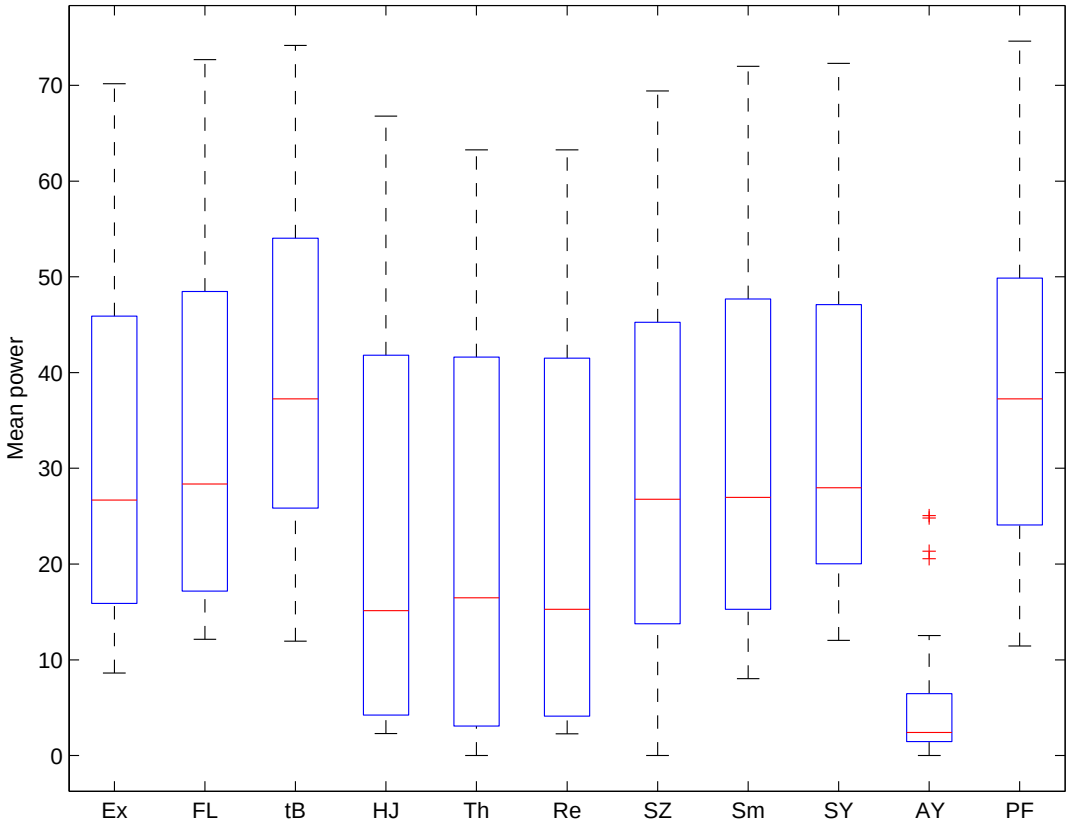


Figure 2.4: Boxplot showing the distribution of $100\,\text{mean}(\alpha) : \alpha' \leq 5\%$ under the alternative hypothesis, over the 30 scenarios in the rows of table 2.8.

Considering Smith's method, it frequently out-performs its simpler variant Shuffle-Z (by quite dramatic amounts for $n = 6$ and discrete interest) in cases for which its average size is valid, suggesting that it should generally be chosen in preference. Compared to FL, however, Sm is generally less powerful for $n = 6$, notably so for discrete nuisance, and most of all, for discrete nuisance and discrete interest. However, these differences do not persist at higher n, where most of the methods converge in terms of power. There are several important exceptions to this trend though. The Adjust-Y method performs extremely badly for larger n, providing empirical backing for Kennedy's assertion that it would be expected to show low power [9]. The parametric F-test is generally less powerful than the permutation methods under cubed-exponential errors for $n = 16$, although, surprisingly, it was slightly more powerful for the same error distribution with $n = 6$.

Disappointingly, the transformed-residual strategies, which showed conservative control of size, also exhibit generally low power. It is not surprising that they fair badly for the extreme situations of $n = 6$ with discrete interest. However, they are also among the least powerful methods for $n = 9$ (with both univariate and multivariate data), including the cases with continuous designs, for which $(n-3)! = 720$ reduced-space permutations are available. Although this is still significantly fewer than the 5000 used for the traditional methods, the rationale of reduced-space permutation is that it includes all of the meaningful permutations [50]. There is limited evidence that one of these approaches (Th) is more powerful than FL at $n = 16$, though its advantages are small and disappear under purely

discrete nuisance. Further investigation would be useful to characterise these differences more carefully, but at this stage, we are forced to conclude that the appealing theoretical basis of Huh and Jhun's suggestion [50] seems to provide it with conservative control of false positives, but lower power than theoretically inexact methods, even in cases where the latter still maintain their size.

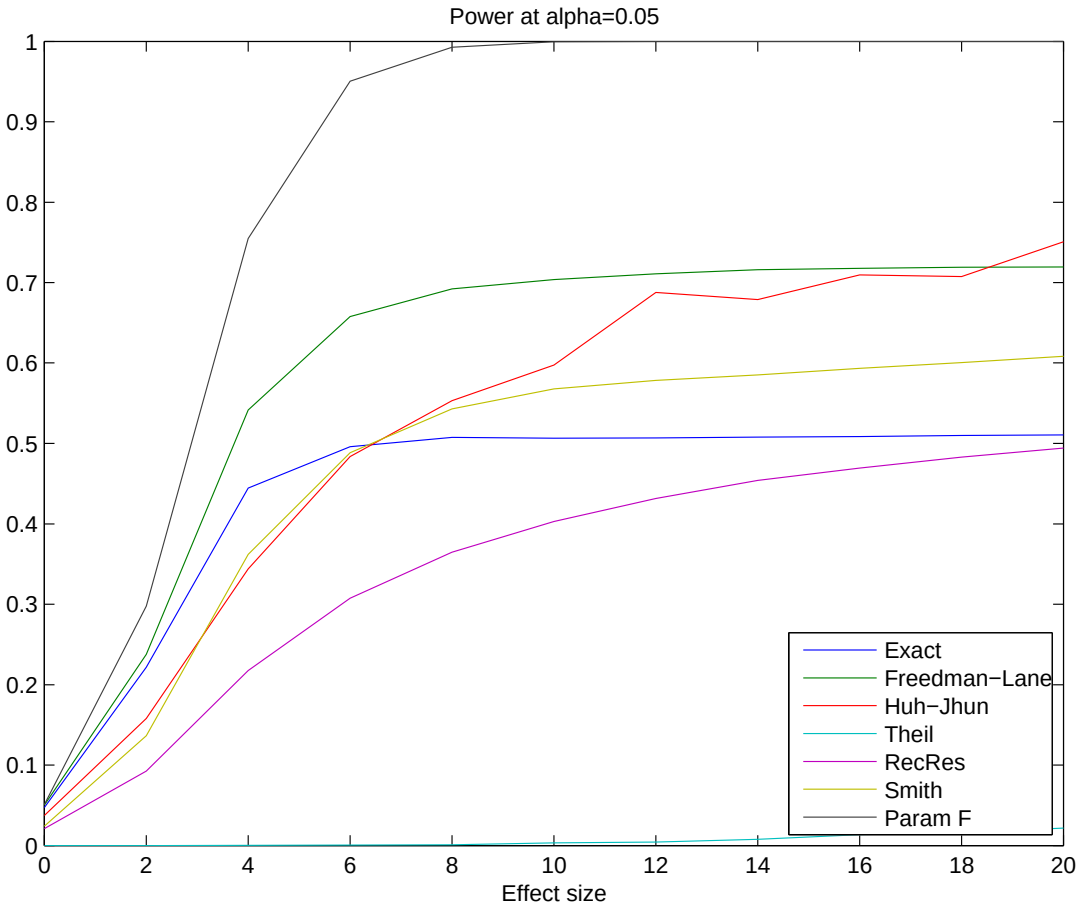


Figure 2.5: Power curve, at $\alpha' = 5\%$, for a subset of methods over a wide range of effect sizes. Discrete interest and nuisance, with $n = 6$ and normal errors.

The extreme case of $n = 6$ with normal errors but discrete interest and nuisance is particularly interesting for two reasons: FL beats the exact method, and the transformed-residual approaches perform particularly badly compared to Ex. To investigate this case more closely, figure 2.5 plots power curves over a wider range of effect sizes than were averaged over for the other tables and figures. Interestingly, all of the permutation methods seem to level out for higher effect sizes, unlike the parametric test which reaches 100% power for sufficient effect (about $b_1 = 10$). This is understandable given that the low number of effective permutations makes it difficult for small p-values to occur. One important aspect of this figure is that the Huh-Jhun method is poor over the 2–8 range of effects used for the other tables, but actually surpasses Sm (and Ex) for larger effect sizes. Considering a less stringent level of $\alpha' = 0.1$ (not shown), to partially address the issue of the small number of permutations, the Th curve is lifted up to more reasonable powers, and FL and Sm manage to reach 100% power. However, Ex still seems to plateau, reaching a power of

about 84% at an effect size of 6, but barely improving to 85% by $b_1 = 20$. Again, we must admit that this case is unrealistically challenging,[36] but it nevertheless seems interesting, and is probably worthy of further exploration.

Table 2.9 examines the variability of power. Intuitively, one would expect the results to be similar to the equivalent ones for variability of rejection rate under the null hypothesis, however, the rankings are quite different for the two performance metrics. The key distinction is that for size, the variability was measured in terms of the RMS error of the observed rejection rate around the expected level; for power, the expected rejection rate is unknown, so we have elected to measure the standard deviation over the range of levels considered. The most obvious effect of this is that the variabilities are much higher than in table 2.5. More importantly though, this lack of ground truth leads to a misleading effect, that lower powers are typically less variable. In particular, the apparent preference for SZ to Sm in terms of power variability (in contrast to the other metrics, which favour Sm) can be seen to arise chiefly from the very low variability that coincides with SZ's very low power for discrete interest covariates. Similarly, the fact that Adjust-Y has the least variable power is meaningless, since it also has the lowest by a large margin. The most interesting result from the table is that ter Braak's method has one of the lowest variabilities, despite having some of the largest average powers. However, the importance of this is limited, in light of the earlier observation that this method's high power seems to stem from poor control of size.

The correlations of the sets of N_s p-values between the exact and the other methods are given in tables 2.10 and 2.11, under the null and alternative hypothesis respectively. The finding that FL has the closest correlation with the exact method under H_0 is unsurprising; however, the closer correlation for Sm under H_1 is interesting, as it implies that although the method is less directly approximating the theory of the exact method compared to FL, Smith's method is still a very good approximation in practice. It is initially counter-intuitive that Kennedy appears to be very strongly correlated with the exact test, in contrast to the investigations of accuracy. However, the likely explanation for this is that the correlation does not penalise a consistent bias; Kennedy's p-values are biased downward, which leads to an inflated false-positive rate, but does not mean that they cannot have a very similar trend over the different simulations. This fact may also detract from the apparently excellent performance of SY under the null hypothesis. In conclusion, correlation with the exact test's p-values is a potentially misleading metric; more weight should be placed on the accuracies and powers in the earlier tables.

2.5.2 Correlations among methods' statistics

The previous section compared the different permutation methods in terms of their size and power, and their p-value correlations over repeated simulations. It is also of interest to know how they compare on an individual simulation in terms of the values of their statistics for each permutation. Anderson and Robinson [40] derived theoretical values for the expected asymptotic (Pearson) correlation between the sets of permutation statistics

[36] With continuous interest and design (not shown) none of the methods' power curves flatten out.

n	m	$\mathcal{E}$	X	Ex	FL	tB	HJ	Th	Re	SZ	Sm	SY	AY	PF
6	1	z	CC	12.38^{9}	12.68^{11}	10.95^{4}	9.443^{1}	9.713^{3}	9.664^{2}	12.22^{8}	12.62^{10}	12.09^{7}	10.96^{5}	12.07^{6}
6	1	z	DC	13.59^{5}	17.25^{9}	11.41^{4}	14.87^{7}	6.87^{3}	14.49^{6}	0.004743^{2}	18.47^{11}	17.85^{10}	0^{1}	16.01^{8}
6	1	z	CD	12.36^{10}	12.15^{8}	10.8^{4}	9.312^{2}	9.387^{3}	9.272^{1}	12.33^{9}	12.05^{7}	12.36^{11}	11.19^{5}	12.03^{6}
6	1	z	DD	12.45^{6}	18.93^{11}	11.67^{5}	14.41^{8}	0.01498^{3}	9.198^{4}	0.006008^{2}	12.64^{7}	15.71^{9}	0^{1}	16.25^{10}
6	1	t_5	CC	10.53^{9}	10.84^{11}	9.37^{4}	7.983^{1}	8.161^{2}	8.183^{3}	10.37^{8}	10.71^{10}	10.29^{6}	9.706^{5}	10.32^{7}
6	1	t_5	DC	12.68^{6}	15.7^{9}	10.51^{4}	12.62^{5}	6.089^{3}	12.89^{7}	0.004427^{2}	16.48^{11}	16.04^{10}	0^{1}	14.96^{8}
6	1	t_5	CD	10.36^{10}	10.18^{6}	9.108^{4}	7.983^{3}	7.977^{2}	7.776^{1}	10.33^{9}	10.21^{7}	10.47^{11}	9.762^{5}	10.27^{8}
6	1	t_5	DD	11.44^{7}	17^{11}	10.7^{5}	12.75^{8}	0.01218^{3}	7.958^{4}	0.003479^{2}	10.79^{6}	13.99^{9}	0^{1}	15.38^{10}
6	1	e^3	CC	5.004^{5}	5.045^{8}	4.474^{2}	5.172^{9}	6.4^{11}	5.211^{10}	5.043^{7}	4.935^{4}	4.603^{3}	5.041^{6}	4.346^{1}
6	1	e^3	DC	6.007^{6}	6.974^{8}	5.64^{5}	7.413^{11}	5.636^{4}	7.199^{9}	0.01629^{2}	7.334^{10}	6.886^{7}	0^{1}	5.328^{3}
6	1	e^3	CD	5.183^{9}	4.68^{3}	4.313^{1}	5.075^{7}	5.826^{11}	5.213^{10}	4.918^{5}	5.022^{6}	5.103^{8}	4.829^{4}	4.337^{2}
6	1	e^3	DD	6.259^{8}	8.269^{11}	5.123^{5}	7.008^{9}	0.2482^{3}	4.842^{4}	0.0215^{2}	5.672^{7}	8.063^{10}	0^{1}	5.461^{6}
9	1	z	CC	12.91^{6}	13.2^{8}	12.65^{2}	13.69^{10}	13.8^{11}	13.59^{9}	12.9^{5}	13.03^{7}	12.88^{4}	1.047^{1}	12.82^{3}
9	1	z	DC	15.46^{8}	14.06^{4}	13.27^{2}	16.53^{11}	16.5^{10}	16.47^{9}	15.04^{7}	14.22^{5}	14.6^{6}	1.019^{1}	13.76^{3}
9	1	z	CM	13.69^{6}	13.84^{8}	13.3^{2}	15.2^{9}	15.24^{10}	15.59^{11}	13.65^{5}	13.81^{7}	13.64^{4}	0.8987^{1}	13.46^{3}
9	1	z	DM	16.1^{8}	13.31^{4}	11.22^{2}	17.82^{9}	18.41^{11}	18.31^{10}	15.82^{7}	13.4^{5}	14.08^{6}	3.209^{1}	11.83^{3}
9	1	z	CD	13.88^{4}	14.03^{8}	13.66^{2}	15.6^{10}	15.42^{9}	15.66^{11}	13.93^{6}	14.01^{7}	13.89^{5}	1.129^{1}	13.77^{3}
9	1	z	DD	16.04^{8}	13.08^{4}	11.63^{2}	18.49^{10}	20.36^{11}	18.37^{9}	15.88^{7}	13.84^{5}	14.31^{6}	1.997^{1}	11.73^{3}
9	2	z	CC	13.23^{9}	13.25^{10}	13.11^{3}	13.33^{11}	13.23^{8}	12.97^{2}	13.16^{4}	13.22^{6}	13.22^{7}	1.326^{1}	13.2^{5}
9	2	z	DC	16.13^{10}	15.41^{3}	15.22^{2}	16.21^{11}	16.1^{9}	15.95^{8}	15.63^{6}	15.49^{5}	15.64^{7}	1.27^{1}	15.41^{4}
9	2	z	CM	14.58^{8}	14.58^{6}	14.43^{2}	14.84^{9}	14.99^{11}	14.98^{10}	14.57^{5}	14.58^{7}	14.56^{4}	1.245^{1}	14.52^{3}
9	2	z	DM	17.62^{8}	15.64^{4}	15.52^{3}	18.21^{9}	18.38^{10}	18.53^{11}	16.95^{7}	15.75^{5}	16.04^{6}	1.782^{1}	15.18^{2}
9	2	z	CD	14.77^{5}	14.8^{7}	14.68^{2}	15.08^{11}	14.93^{9}	14.97^{10}	14.81^{8}	14.79^{6}	14.77^{4}	1.373^{1}	14.76^{3}
9	2	z	DD	17.98^{8}	15.92^{3}	16.29^{5}	18.48^{9}	20.21^{11}	18.6^{10}	17.4^{7}	16.27^{4}	16.53^{6}	1.479^{1}	15.28^{2}
16	1	e^3	CC	3.563^{9}	3.539^{7}	3.444^{5}	3.433^{4}	3.56^{8}	3.38^{2}	3.805^{11}	3.519^{6}	3.571^{10}	1.489^{1}	3.4^{3}
16	1	e^3	DC	4.351^{10}	4.338^{9}	4.18^{5}	4.125^{4}	4.284^{7}	4.039^{2}	4.794^{11}	4.275^{6}	4.33^{8}	1.637^{1}	4.113^{3}
16	1	e^3	CM	4.162^{6}	4.206^{9}	4.02^{5}	3.966^{4}	4.19^{8}	3.91^{2}	4.407^{11}	4.288^{10}	4.166^{7}	1.592^{1}	3.963^{3}
16	1	e^3	DM	5.36^{7}	5.438^{8}	5.183^{3}	5.217^{5}	5.448^{9}	5.193^{4}	5.758^{11}	5.614^{10}	5.356^{6}	3.821^{1}	5.114^{2}
16	1	e^3	CD	4.211^{9}	4.179^{7}	3.93^{3}	3.961^{4}	4.076^{5}	4.077^{6}	4.276^{11}	4.234^{10}	4.202^{8}	1.556^{1}	3.9^{2}
16	1	e^3	DD	5.606^{10}	5.544^{7}	5.168^{2}	5.352^{4}	5.462^{5}	5.462^{6}	5.693^{11}	5.555^{8}	5.604^{9}	3.972^{1}	5.219^{3}
				7.633	7.4	3.3	7.167	7.1	6.433	6.6	7.167	7.133	1.8	4.267

Table 2.9: Power variability, quantified by $100\,\mathrm{std}(\alpha) : \alpha' \leq 5\%$, smaller is better. Superscripts show ranks, with 1 = best. The final row shows the mean rank for each method.

n	m	$\mathcal{E}$	X	FL	tB	HJ	Th	Re	SZ	Sm	SY	AY	Ke	PF
6	1	z	CC	0.994^{2}	0.991^{6}	0.899^{9}	0.891^{11}	0.899^{10}	0.993^{4}	0.995^{1}	0.983^{7}	0.964^{8}	0.992^{5}	0.994^{3}
6	1	z	DC	0.971^{2}	0.95^{5}	0.866^{10}	0.77^{11}	0.898^{9}	0.947^{6}	0.968^{4}	0.931^{7}	0.899^{8}	0.969^{3}	0.974^{1}
6	1	z	CD	0.994^{2}	0.991^{7}	0.896^{10}	0.9^{9}	0.885^{11}	0.992^{4}	0.993^{3}	0.992^{5}	0.95^{8}	0.991^{6}	0.994^{1}
6	1	z	DD	0.972^{3}	0.961^{6}	0.874^{9}	0.844^{11}	0.868^{10}	0.944^{7}	0.975^{2}	0.969^{5}	0.929^{8}	0.971^{4}	0.98^{1}
6	1	t_5	CC	0.994^{1}	0.991^{6}	0.907^{9}	0.89^{11}	0.898^{10}	0.992^{4}	0.994^{2}	0.984^{7}	0.963^{8}	0.992^{5}	0.993^{3}
6	1	t_5	DC	0.972^{2}	0.952^{5}	0.873^{10}	0.773^{11}	0.897^{9}	0.944^{6}	0.968^{4}	0.936^{7}	0.9^{8}	0.97^{3}	0.972^{1}
6	1	t_5	CD	0.993^{1}	0.99^{7}	0.897^{10}	0.901^{9}	0.887^{11}	0.991^{5}	0.993^{2}	0.993^{4}	0.945^{8}	0.99^{6}	0.993^{3}
6	1	t_5	DD	0.971^{3}	0.959^{6}	0.882^{9}	0.836^{11}	0.867^{10}	0.941^{7}	0.973^{2}	0.969^{5}	0.924^{8}	0.97^{4}	0.977^{1}
6	1	e^3	CC	0.981^{1}	0.971^{4}	0.868^{10}	0.823^{11}	0.873^{9}	0.968^{6}	0.975^{3}	0.979^{2}	0.934^{8}	0.97^{5}	0.963^{7}
6	1	e^3	DC	0.941^{1}	0.898^{6}	0.794^{10}	0.715^{11}	0.801^{9}	0.861^{7}	0.931^{3}	0.915^{4}	0.81^{8}	0.932^{2}	0.901^{5}
6	1	e^3	CD	0.972^{2}	0.959^{6}	0.861^{10}	0.87^{9}	0.86^{11}	0.959^{5}	0.969^{3}	0.99^{1}	0.909^{8}	0.967^{4}	0.956^{7}
6	1	e^3	DD	0.923^{3}	0.888^{6}	0.803^{10}	0.796^{11}	0.808^{9}	0.888^{7}	0.931^{2}	0.959^{1}	0.867^{8}	0.922^{4}	0.891^{5}
9	1	z	CC	0.999^{2}	0.999^{4}	0.964^{10}	0.968^{8}	0.968^{9}	0.999^{5}	0.999^{3}	0.998^{6}	0.681^{11}	0.989^{7}	0.999^{1}
9	1	z	DC	0.998^{2}	0.997^{4}	0.959^{10}	0.97^{8}	0.963^{9}	0.995^{5}	0.998^{3}	0.994^{6}	0.707^{11}	0.988^{7}	0.998^{1}
9	1	z	CM	0.999^{2}	0.999^{5}	0.966^{9}	0.965^{10}	0.968^{8}	0.999^{4}	0.999^{3}	0.997^{6}	0.727^{11}	0.988^{7}	0.999^{1}
9	1	z	DM	0.998^{3}	0.996^{4}	0.967^{9}	0.966^{10}	0.97^{8}	0.995^{5}	0.998^{2}	0.993^{6}	0.813^{11}	0.988^{7}	0.998^{1}
9	1	z	CD	0.999^{2}	0.999^{6}	0.968^{8}	0.953^{10}	0.957^{9}	0.999^{5}	0.999^{3}	0.999^{4}	0.726^{11}	0.989^{7}	0.999^{1}
9	1	z	DD	0.998^{2}	0.997^{5}	0.967^{8}	0.949^{10}	0.953^{9}	0.995^{6}	0.998^{3}	0.998^{4}	0.834^{11}	0.988^{7}	0.998^{1}
9	2	z	CC	0.999^{2}	0.999^{5}	0.968^{10}	0.972^{7}	0.969^{8}	0.999^{4}	0.999^{3}	0.999^{6}	0.577^{11}	0.969^{9}	0.999^{1}
9	2	z	DC	0.999^{2}	0.998^{4}	0.962^{10}	0.974^{7}	0.966^{9}	0.997^{6}	0.998^{3}	0.997^{5}	0.604^{11}	0.969^{8}	0.999^{1}
9	2	z	CM	0.999^{2}	0.999^{5}	0.967^{10}	0.967^{9}	0.971^{7}	0.999^{4}	0.999^{3}	0.999^{6}	0.603^{11}	0.968^{8}	0.999^{1}
9	2	z	DM	0.998^{3}	0.997^{5}	0.967^{10}	0.969^{8}	0.974^{7}	0.997^{6}	0.998^{2}	0.997^{4}	0.701^{11}	0.968^{9}	0.999^{1}
9	2	z	CD	0.999^{3}	0.999^{6}	0.971^{7}	0.957^{10}	0.961^{9}	0.999^{5}	0.999^{4}	0.999^{2}	0.609^{11}	0.968^{8}	0.999^{1}
9	2	z	DD	0.998^{3}	0.998^{5}	0.969^{7}	0.955^{10}	0.958^{9}	0.997^{6}	0.998^{4}	0.999^{2}	0.724^{11}	0.968^{8}	0.999^{1}
16	1	e^3	CC	0.998^{2}	0.996^{4}	0.987^{8}	0.978^{10}	0.98^{9}	0.992^{6}	0.997^{3}	1^{1}	0.728^{11}	0.994^{5}	0.992^{7}
16	1	e^3	DC	0.997^{2}	0.994^{4}	0.977^{8}	0.972^{10}	0.974^{9}	0.982^{7}	0.996^{3}	0.999^{1}	0.71^{11}	0.993^{5}	0.985^{6}
16	1	e^3	CM	0.998^{2}	0.996^{4}	0.984^{8}	0.979^{10}	0.98^{9}	0.996^{5}	0.997^{3}	0.999^{1}	0.753^{11}	0.994^{6}	0.991^{7}
16	1	e^3	DM	0.995^{2}	0.991^{4}	0.969^{9}	0.97^{8}	0.957^{10}	0.99^{5}	0.995^{3}	0.998^{1}	0.822^{11}	0.99^{6}	0.978^{7}
16	1	e^3	CD	0.997^{2}	0.995^{4}	0.983^{9}	0.981^{10}	0.985^{8}	0.993^{6}	0.997^{3}	1^{1}	0.76^{11}	0.994^{5}	0.99^{7}
16	1	e^3	DD	0.994^{2}	0.991^{5}	0.971^{10}	0.977^{9}	0.979^{8}	0.989^{6}	0.994^{3}	0.999^{1}	0.84^{11}	0.991^{4}	0.98^{7}
				2.1	5.1	9.2	9.667	9.067	5.467	2.833	3.933	9.8	5.8	3.033

Table 2.10: Correlations of p-values from each method with those from the exact method, over the 10000 simulations, under the null hypothesis. Superscripts show ranks, with 1 = best. The final row shows the mean rank for each method.

n	m	$\mathcal{E}$	X	FL	tB	HJ	Th	Re	SZ	Sm	SY	AY	PF
6	1	z	CC	0.992^{4}	0.985^{6}	0.858^{9}	0.815^{10}	0.86^{8}	0.993^{2}	0.993^{3}	0.987^{5}	0.879^{7}	0.993^{1}
6	1	z	DC	0.86^{6}	0.892^{2}	0.73^{9}	0.388^{10}	0.83^{7}	0.871^{4}	0.874^{3}	0.87^{5}	0.753^{8}	0.921^{1}
6	1	z	CD	0.991^{5}	0.987^{6}	0.831^{9}	0.848^{8}	0.826^{10}	0.991^{4}	0.991^{3}	0.995^{1}	0.861^{7}	0.993^{2}
6	1	z	DD	0.245^{9}	0.427^{5}	0.392^{7}	0.184^{10}	0.396^{6}	0.726^{1}	0.367^{8}	0.624^{3}	0.677^{2}	0.467^{4}
6	1	t_5	CC	0.994^{2}	0.988^{6}	0.89^{9}	0.848^{10}	0.892^{8}	0.994^{4}	0.994^{1}	0.991^{5}	0.92^{7}	0.994^{3}
6	1	t_5	DC	0.892^{4}	0.894^{3}	0.765^{9}	0.406^{10}	0.834^{7}	0.879^{6}	0.899^{2}	0.889^{5}	0.774^{8}	0.934^{1}
6	1	t_5	CD	0.993^{3}	0.989^{6}	0.879^{8}	0.874^{9}	0.851^{10}	0.993^{5}	0.993^{4}	0.996^{1}	0.891^{7}	0.994^{2}
6	1	t_5	DD	0.654^{7}	0.693^{6}	0.564^{8}	0.313^{10}	0.544^{9}	0.802^{2}	0.71^{5}	0.808^{1}	0.743^{4}	0.746^{3}
6	1	e^3	CC	0.99^{2}	0.984^{4}	0.904^{9}	0.879^{10}	0.907^{8}	0.984^{5}	0.987^{3}	0.996^{1}	0.952^{7}	0.981^{6}
6	1	e^3	DC	0.974^{2}	0.962^{5}	0.892^{9}	0.823^{10}	0.901^{8}	0.952^{6}	0.972^{3}	0.981^{1}	0.902^{7}	0.963^{4}
6	1	e^3	CD	0.986^{2}	0.979^{5}	0.925^{8}	0.911^{9}	0.903^{10}	0.981^{4}	0.985^{3}	0.999^{1}	0.932^{7}	0.978^{6}
6	1	e^3	DD	0.957^{3}	0.94^{5}	0.872^{9}	0.838^{10}	0.885^{8}	0.936^{6}	0.962^{2}	0.989^{1}	0.915^{7}	0.943^{4}
9	1	z	CC	0.996^{5}	0.996^{6}	0.92^{9}	0.948^{7}	0.921^{8}	0.996^{3}	0.996^{4}	0.998^{1}	0.308^{10}	0.998^{2}
9	1	z	DC	0.952^{6}	0.961^{4}	0.864^{8}	0.879^{7}	0.862^{9}	0.96^{5}	0.962^{3}	0.984^{1}	0.221^{10}	0.966^{2}
9	1	z	CM	0.996^{4}	0.995^{6}	0.9^{9}	0.949^{8}	0.955^{7}	0.997^{2}	0.996^{5}	0.996^{3}	0.286^{10}	0.998^{1}
9	1	z	DM	0.95^{6}	0.965^{2}	0.916^{8}	0.928^{7}	0.904^{9}	0.952^{5}	0.96^{3}	0.972^{1}	0.336^{10}	0.952^{4}
9	1	z	CD	0.997^{3}	0.995^{6}	0.932^{7}	0.903^{8}	0.889^{9}	0.996^{5}	0.996^{4}	0.998^{1}	0.285^{10}	0.997^{2}
9	1	z	DD	0.86^{5}	0.816^{6}	0.721^{7}	0.681^{8}	0.637^{9}	0.863^{4}	0.887^{2}	0.934^{1}	0.41^{10}	0.868^{3}
9	2	z	CC	0.995^{5}	0.995^{6}	0.917^{8}	0.936^{7}	0.915^{9}	0.995^{4}	0.995^{3}	0.998^{1}	0.26^{10}	0.997^{2}
9	2	z	DC	0.955^{6}	0.962^{4}	0.843^{9}	0.871^{7}	0.862^{8}	0.956^{5}	0.963^{3}	0.983^{1}	0.22^{10}	0.963^{2}
9	2	z	CM	0.995^{4}	0.994^{6}	0.936^{7}	0.934^{8}	0.932^{9}	0.996^{3}	0.995^{5}	0.997^{1}	0.188^{10}	0.997^{2}
9	2	z	DM	0.943^{5}	0.968^{2}	0.884^{9}	0.918^{7}	0.894^{8}	0.951^{4}	0.953^{3}	0.969^{1}	0.262^{10}	0.942^{6}
9	2	z	CD	0.995^{3}	0.994^{6}	0.914^{7}	0.897^{8}	0.886^{9}	0.994^{5}	0.995^{4}	0.998^{1}	0.252^{10}	0.997^{2}
9	2	z	DD	0.9^{4}	0.831^{6}	0.773^{7}	0.77^{8}	0.724^{9}	0.885^{5}	0.916^{2}	0.96^{1}	0.357^{10}	0.912^{3}
16	1	e^3	CC	0.999^{2}	0.998^{4}	0.99^{7}	0.986^{8}	0.986^{9}	0.995^{5}	0.998^{3}	1^{1}	0.687^{10}	0.995^{6}
16	1	e^3	DC	0.998^{2}	0.997^{4}	0.989^{7}	0.984^{9}	0.985^{8}	0.991^{6}	0.998^{3}	1^{1}	0.714^{10}	0.993^{5}
16	1	e^3	CM	0.999^{2}	0.997^{4}	0.99^{7}	0.988^{8}	0.987^{9}	0.997^{5}	0.998^{3}	1^{1}	0.733^{10}	0.995^{6}
16	1	e^3	DM	0.998^{2}	0.997^{4}	0.988^{7}	0.988^{8}	0.983^{9}	0.996^{5}	0.998^{3}	1^{1}	0.878^{10}	0.992^{6}
16	1	e^3	CD	0.999^{2}	0.997^{4}	0.989^{8}	0.989^{9}	0.991^{7}	0.996^{5}	0.998^{3}	1^{1}	0.747^{10}	0.994^{6}
16	1	e^3	DD	0.998^{2}	0.997^{4}	0.987^{9}	0.99^{8}	0.991^{7}	0.996^{5}	0.998^{3}	1^{1}	0.898^{10}	0.993^{6}
				3.9	4.767	8.1	8.533	8.367	4.333	3.3	1.667	8.6	3.433

Table 2.11: Correlations of p-values from each method with those from the exact method, over the 10000 simulations, under the alternative hypothesis (with an effect size of 6). Superscripts show ranks, with 1 = best. The final row shows the mean rank for each method.

from four methods: exact, Freedman-Lane, ter Braak and Shuffle-Y. Their results assume that the permutation statistic is the signed Pearson correlation coefficient and that only one interest- and one nuisance-covariate are present in the model; they can be summarised as follows:

$$\rho(Ex, FL) = 1 \tag{2.34}$$

$$\rho(Ex, tB) = \rho(FL, tB) = \sqrt{1 - r^2} \tag{2.35}$$

$$\rho(Ex, SY) = \rho(FL, SY) = \sqrt{\frac{1}{1 + g^2}} \tag{2.36}$$

$$\rho(tB, SY) = \sqrt{\frac{1 - r^2}{1 + g^2}},$$

where (using the centring matrix $\bar{M} = I - 1_{n\times 1}1^{+}_{n\times 1}$)

$$r = \rho(y, x_1|x_0) = \frac{(R_0 y)^T (R_0 x_1)}{\|R_0 y\| \, \|R_0 x_1\|}$$
$$g = \frac{(\bar{M} x_0)^T (\bar{M} y)}{\|\bar{M} x_0\| \, \|R_0 y\|},$$

and x_0 is the single nuisance-covariate, but R_0 is derived from X_0 including a constant term, i.e. $X_0 = [x_0 \; 1_{n\times 1}]$.

In addition to the theoretical expressions, Anderson and Robinson gave predicted and measured correlations for a single set of simulated data [40].[37] They chose $n = 40$ and $b_1 = b_0 = 1$ (i.e. the alternative hypothesis holds), and sampled individual vectors x_1 and x_0 with IID elements from a uniform distribution on the interval (0, 3). A single vector of errors e with IID standard normal elements was sampled and used to create the data $y = x_1 b_1 + x_0 b_0 + e$. They considered 999 permutations (which we assume did not include the original labelling, so that there were 1000 permutations including the identity).

Anderson and Robinson observed close agreement between the predicted and measured correlations [40], however, their relatively large n and investigation of a single design and single noise-realisation weaken this finding. Here, we repeat their single simulation 100 times, and also produce a further 100 simulations for a more challenging low DF example with $n = 6$.

Only four permutation methods were studied in [40]. It is reasonable to exclude Kennedy's method and Adjust-Y, given their unacceptable performance in terms of size and power respectively. It is also impossible to compare the sets of statistics under permutation from the three transformed-residual permutation strategies to those from the other methods, since the transformation leads to a reduced number of permutations. (It is, however, possible to compare the transformed-residual methods to each other, and this will be done below.) After the above exclusions there are six methods remaining; in addition to the four analysed in [40], the results here include Shuffle-Z and Smith's

[37] Note that the predicted correlations between methods other than Ex and FL are data-dependent because r and g involve y.

method.

It seems plausible that the behaviour in practice (as opposed to the asymptotic expectations) of the different methods could depend on whether the simulated model has non-zero values of the true nuisance- and interest-parameters (as simulated in [40]), or one of the other three combinations: non-zero nuisance, but null hypothesis holds; zero nuisance and non-zero interest; both nuisance and interest zero. In particular, it is obvious that Shuffle-Y is equivalent to the exact method if the (unobservable) true nuisance-parameters are zero, and the difference between Freedman-Lane and ter Braak's method concerns the choice of reduced or full model residuals. Figures presented here encompass all four combinations of zero or non-zero interest and nuisance.

Like the t-statistic, the signed Pearson correlation is limited to single covariates, so it is also of interest to compute correlations between sets of permutation statistics using ρ^2, which can be generalised to the squared coefficient of determination for designs with multiple covariates.

Results

Table 2.12 presents the mean correlations among the six methods, extending the case considered by Anderson and Robinson ($n = 40$ and $b_1 = b_0 = 1$). In each case the correlations are between the 999 values of the statistic under the non-identity permutations for a pair of methods. Both ρ and ρ^2 are considered as the test statistic. We can observe that for this case, the correlations based on ρ^2 are broadly similar to those based on ρ; for example, the relative rankings in terms of correlations between the exact and other methods is the same for both test statistics. Results based on the t- and F-statistic for this example were very similar, and are not shown here.

	Ex	FL	tB	SZ	Sm	SY
Ex	1	0.9786	0.5249	0.9565	0.9772	0.6917
FL	0.9895	1	0.5369	0.9769	0.9981	0.6871
tB	0.7306	0.7383	1	0.5248	0.5362	0.3634
SZ	0.9782	0.9887	0.7304	1	0.9781	0.6797
Sm	0.9892	0.9997	0.7382	0.989	1	0.6939
SY	0.8356	0.8312	0.6137	0.8219	0.8309	1

Table 2.12: Average Pearson correlations among the different permutation methods' sets of statistics. Both interest and nuisance-parameters were non-zero. Each permutation set included 999 randomly sampled non-trivial permutations, and the correlations were averaged over 100 repeated simulations (with different designs and noise realisations). The correlations below the diagonal are based on the use of signed correlation coefficient as the test statistic (as in [40]); above the diagonal, equivalent results based on the squared correlation coefficient are given.

The averages over the 100 simulations of the theoretically predicted correlations using ρ are as follows: Ex and FL, 1; Ex and tB, 0.739; Ex and SY, 0.8306; tB and SY, 0.6131. There is excellent agreement between the predicted and measured results, as reported in [40]. We observe high correlation of FL with Ex, slightly lower correlation of SY with Ex or FL, and lower still between tB and FL or Ex. Unsurprisingly, we find that SZ and Sm

are very closely related, however, much more interesting is the new result that Smith's method is very strongly correlated with FL; the correlation between Sm and Ex is virtually identical to that of FL and Ex, and is notably higher than the correlations of the other methods with the exact one.

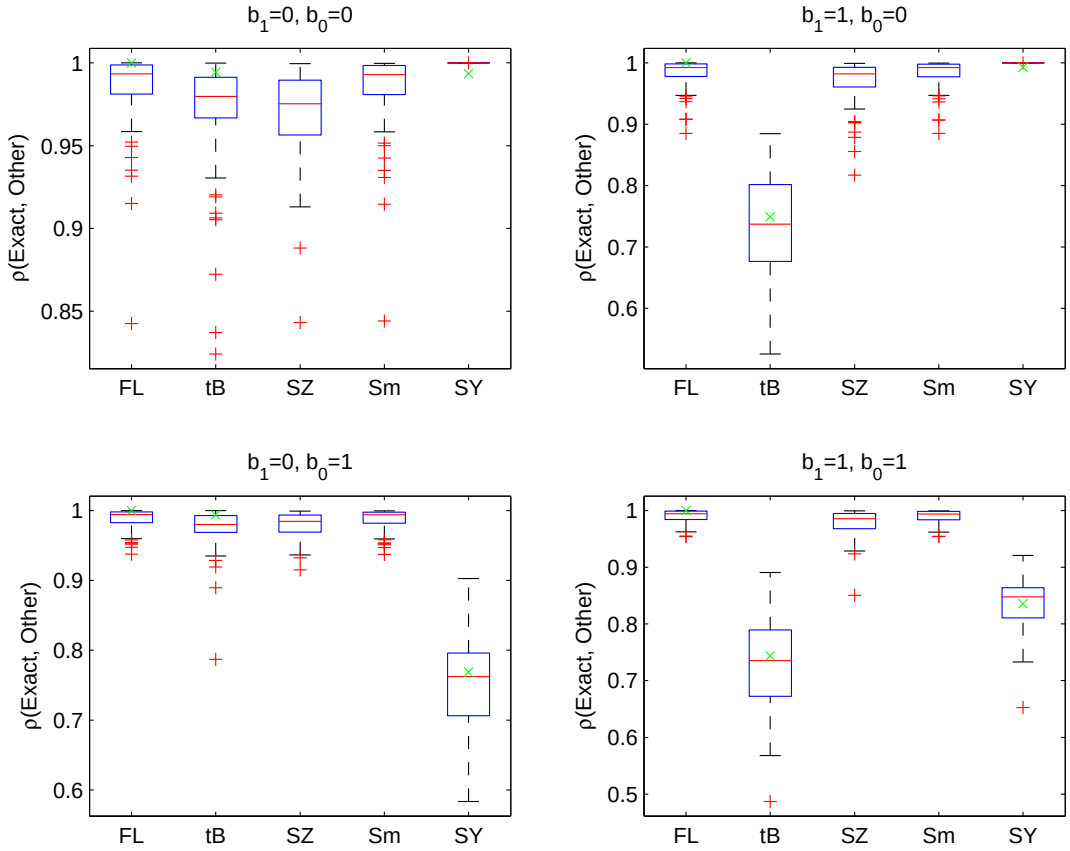


Figure 2.6: Boxplot showing the distribution of correlations between the exact and other methods, over 100 simulations, with $n = 40$, and using ρ as the test statistic. The medians of the three theoretical results from equations (2.34), (2.35) and (2.36) are plotted as green crosses.

Figure 2.6 focusses solely on the correlations of the five practical methods with the hypothetical exact method (and not with each other), and shows (in the bottom-right panel) the distribution over the 100 simulations corresponding to the average shown in the first column of table 2.12. The other three panels show the equivalent results for different values of the interest- and nuisance-parameters. As noted above, SY is identical to Ex if there is truly no nuisance (top panels); it seems to differ most from Ex when there is non-zero nuisance but zero interest. The difference between tB and Ex is greater when the alternative hypothesis holds; however, this difference should not be interpreted as a failing of ter Braak's method, since it is designed to use the estimated interest-parameter in an attempt to reduce the variance under permutation of the test statistic [8], i.e. it is not attempting to reproduce the exact method, which assumes the null hypothesis holds.[38]

[38] Arguably, if the alternative hypothesis is true $Y - X_0 B_0 = E + X_1 B_1$ will not generally be exchangeable; however, in common with parametric tests, Anderson and Robinson's exact method assumes that the null hypothesis is true. Note that this means it ignores a truly non-zero interest-parameter, even though it uses

In all four cases, there is good agreement between the theoretically predicted and observed correlations. Interestingly, Smith's method (for which the theoretical correlation with the exact method is unknown) is barely distinguishable from FL.

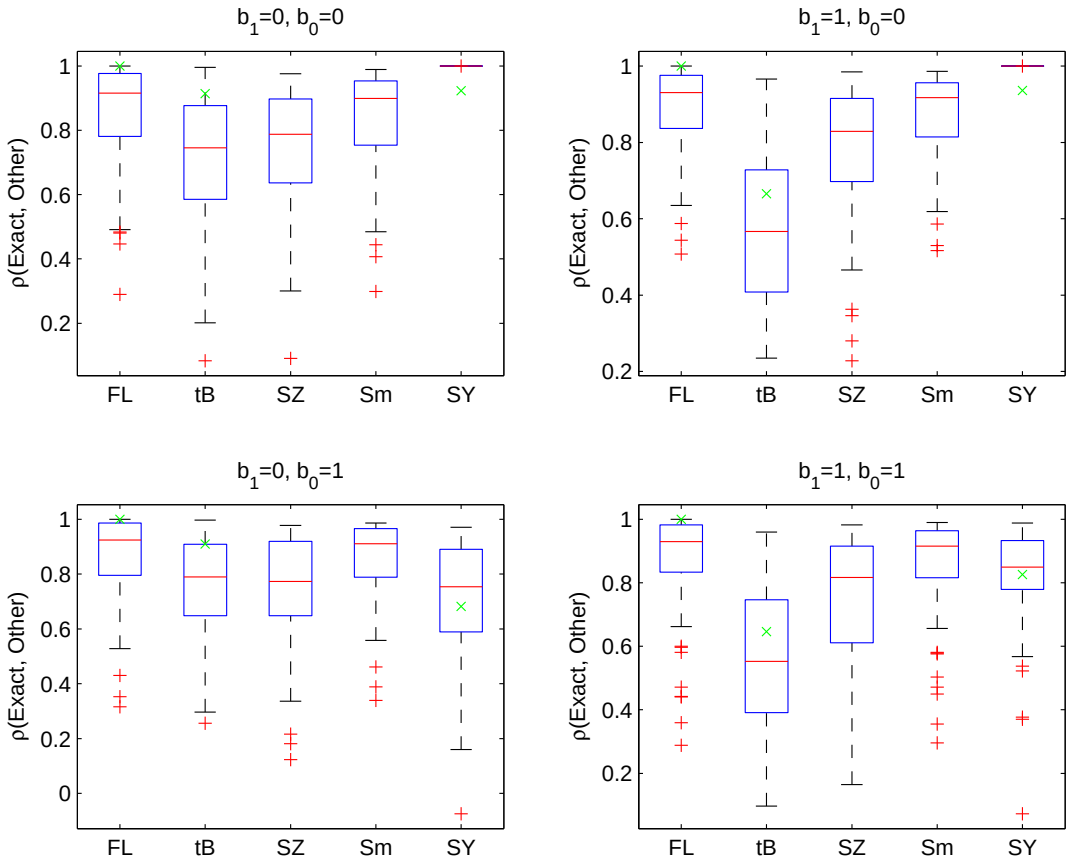


Figure 2.7: Boxplot of correlations, as in figure 2.6, but with $n = 6$, and all 719 non-trivial permutations used to compute the correlations.

Equivalent results for 100 simulations with $n = 6$, using the exhaustive set of $6! - 1 = 719$ non-identity permutations, are shown in figure 2.7. As expected, the correlations are much more variable, with some low values being observed, even for FL and Sm. The theoretical values occasionally fall outside of the interquartile range, but are still indicative of the relative ordering of the results. Compared to the $n = 40$ results, Shuffle-Y now seems closer to the other methods in the cases of non-zero nuisance. Again, Smith's method performs very similarly to FL.

Permutation-based p-values will be equivalent for ρ and t, and for ρ^2 and F, thanks to their monotonic relationships. However, the Pearson correlations based on these pairs of statistics will not be equal (though their Spearman rank correlations obviously would be), so it is potentially of interest to compare the permutation testing methods using these statistics. The equivalent of figure 2.6, with $n = 40$, but using ρ^2 (not shown) exhibits a very similar pattern among the permutation methods and the four classes of simulation, but with generally reduced values of the correlations. Similarly, the t-statistic produces very similar results to ρ, and the F-statistic appears very similar to ρ^2. However, in the

the unobservable true nuisance-parameter.

$n = 6$ simulations, these differences become much more pronounced. Figure 2.8 illustrates the results using the F-statistic (equivalent for this single-interest design to t^2). The correlations are now dramatically lower and more variable; negative correlations can now be observed for all of the methods in some of the simulations. It is therefore important to note that despite FL's perfect asymptotic correlation with the exact method (using the correlation coefficient as a test statistic, for a single interest-parameter), in practice, with small data-sets and tests of multiple interest covariates, one cannot expect the statistics from FL (or any of the other methods) to closely match those from the hypothetical exact test.

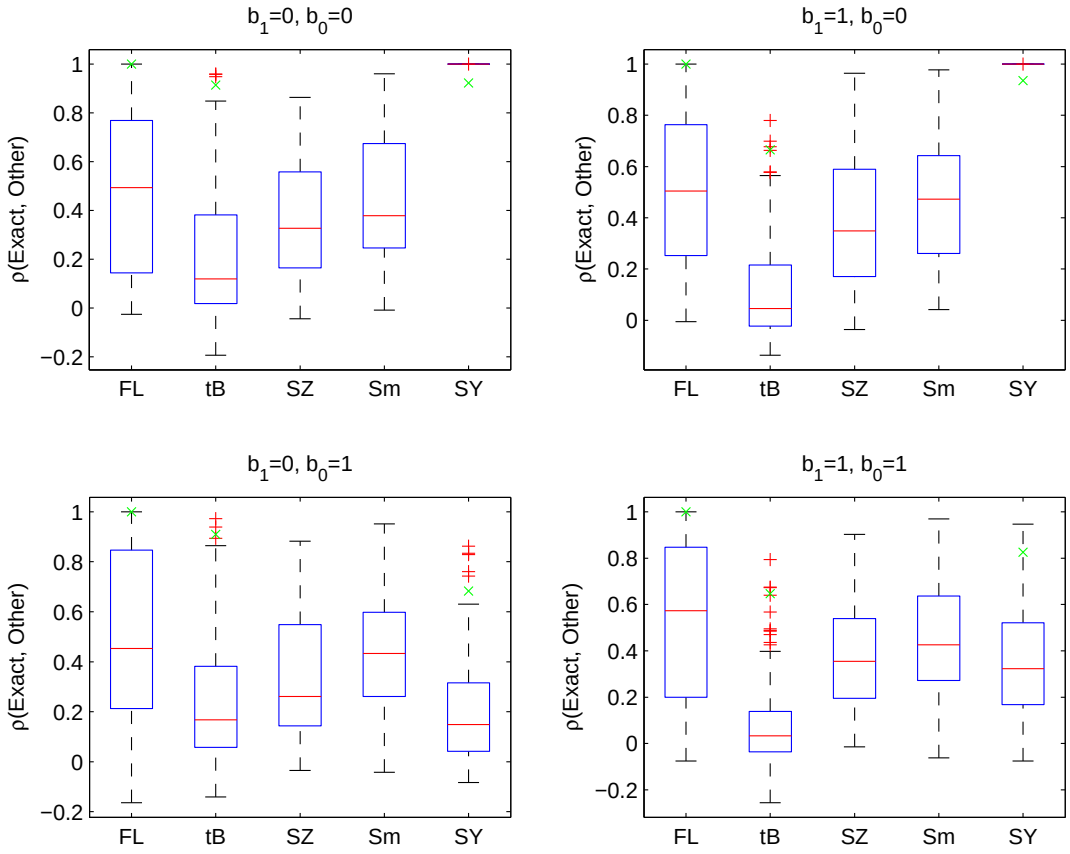


Figure 2.8: Boxplot of correlations, for $n = 6$, as in figure 2.7, but based on the F-statistic.

Transformed-residual permutation strategies are compared in figure 2.9. It is clear that there is very little correlation between the sets of statistics under permutation for the Huh-Jhun method and those for either of the other two approaches. It may seem surprising that there is much less similarity between two versions of USU^TY using different U matrices (which satisfy the same key properties, $UU^T = R$, etc.) than there is between two versions of $RSRY$ using different R matrices (which correspond to distinct models, i.e. R_0 for FL and R for tB). However, in the case of FL and tB, the same permutations are carried out, while the transformed-residual strategies effectively perform different permutations with respect to the original data, since different combinations of the rows are transformed into the reduced permutation space. This might also explain the observation that Th and Re can show greater correlation in some simulations. Both the optimal BLUS residuals and

the recursive residuals are derived from simple selection matrices, which presumably can result in more similar effective permutations than the U matrix from the SVD in Huh and Jhun's method, which is likely to be the BLUS estimate for a more complicated M_L matrix (2.29). Note that Re is not BLUS for its selection matrix, but it might nevertheless be expected that its M_L will be closer to the class of selection matrices than the more random mix implied by HJ. However, since any matrix U_L leads to a class of non-unique M_L, it is difficult to prove this speculative explanation of the observed results. In any case, the basic conclusion is that for a particular permutation within a particular design, the three different transformed residual methods can give very different statistics.

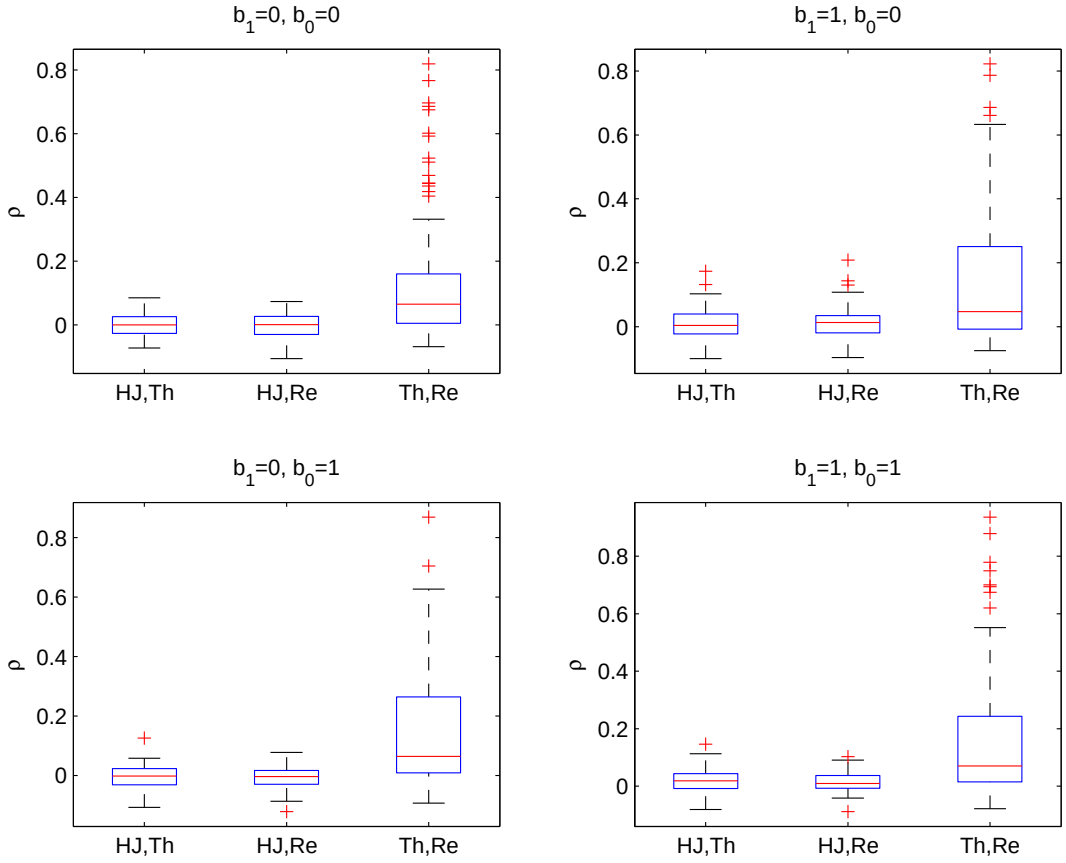


Figure 2.9: Correlations among transformed-residual permutation strategies, using the signed correlation coefficient as the test statistic. 100 simulations with $n = 40$, and 999 non-trivial permutations were performed.

2.5.3 P-value precision

If the exhaustive set of permutations produces a p-value of p_e, the estimated p-value from a random subset of N_p permutations (not including the original) may be assumed to be approximately normally distributed with mean p_e and variance $p_e(1 - p_e)/N_p$ [68]. Table 2.13 uses this expression to evaluate 95% confidence intervals for common values of p_e and numbers of permutations ranging from 500 to 20,000. For example, for the half-width of the confidence interval to be less than 10% of the nominal p-value for $p_e = 0.05$, we find approximately 7,300 permutations are required. A common value in use in neuroimaging

is 5000.

	$p = 1\%$		$p = 5\%$		$p = 10\%$	
N_p	Lower	Upper	Lower	Upper	Lower	Upper
20000	0.8621	1.138	4.698	5.302	9.584	10.42
10000	0.805	1.195	4.573	5.427	9.412	10.59
5000	0.7242	1.276	4.396	5.604	9.168	10.83
2000	0.5639	1.436	4.045	5.955	8.685	11.31
1000	0.3833	1.617	3.649	6.351	8.141	11.86
500	0.1279	1.872	3.09	6.91	7.37	12.63

Table 2.13: Theoretical 95% confidence limits for p-values from exhaustive permutation, $p_e \in \{0.01, 0.05, 0.1\}$, for various numbers of permutations N_p.

Edgington's theory is very frequently cited in recent literature (e.g. [1]), but seems not to have been evaluated in practice to the same extent. Therefore, we briefly explore the precision of p-values from random sampling of the set of permutations, compared to exhaustive evaluation. We simulated a random $\mathrm{N}(0, 1)$ independent vector x for $n = 9$, and a dependent variable $y = xf + e$ where e was a second $\mathrm{N}(0, 1)$ vector, and f was a measure of effect size. We investigated two effect sizes, $f = 0$ to simulate a true null hypothesis, and a value of f chosen to give a parametric p-value for the correlation between x and y equal to 0.05.[39] (The parametric p-value for the null case was 0.3342.)

For $n = 9$, there are $9! = 362,880$ possible permutations. We evaluated the squared correlation coefficient for this complete set, and then considered various subsets randomly sampled from it. We chose to sample without replacement, though it probably makes little difference, meaning 18 different 20,000 permutation subsets were available. Figure 2.10 shows the results under the null hypothesis, and figure 2.11 the results for the parametric p-value of 0.05. The exhaustive p-values were estimated as 0.3335 and 0.0507 respectively, and in both cases there was very good agreement between the theoretical and empirical intervals.

We conclude this section by noting that Anderson and Robinson [40] use the same expression to estimate a confidence interval for the type I error over $N_s = 10000$ simulations, each with $N_p = 1000$ permutations. In this case, the uncertainty in the individual p-values is not directly relevant, and the confidence interval is based solely on N_s. This approach was also used earlier in this chapter, for example in table 2.3.

2.5.4 Class of permutation sampling

With basic models such as simple regression or one-way ANOVA, the complete set of permutations from which one may randomly sample has a straightforward and unambiguous form. For example, simple regression has available the $n!$ different orders of the data (or covariate). With a categorical covariate, the set of useful permutations is reduced by the following fact: shuffling data in such a way that it remains paired with equal values of the covariate will lead to unhelpful duplication of statistics. Similarly, in ANOVA, the n_l

[39] This was achieved by optimising $(p(y(f), x; e) - 0.05)^2$ as a function of f with fixed e using standard MATLAB routines.

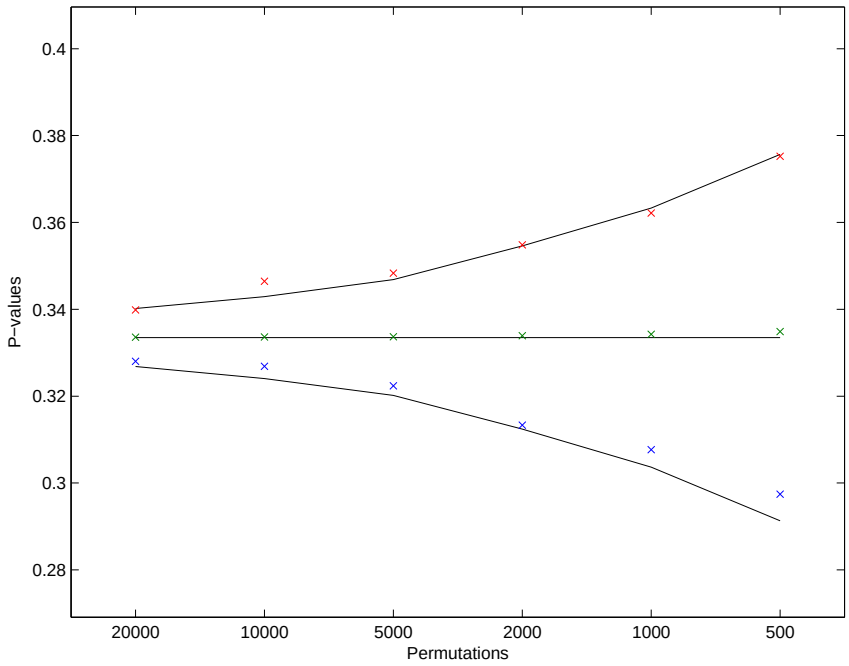


Figure 2.10: P-value precision as a function of number of permutations, under a true null hypothesis. The black lines show the exhaustive p-value and its upper and lower limits for a 95% confidence interval. The green, blue and red crosses show respectively: the mean of the subsets' estimated p-values, and their 2.5 and 97.5 percentiles.

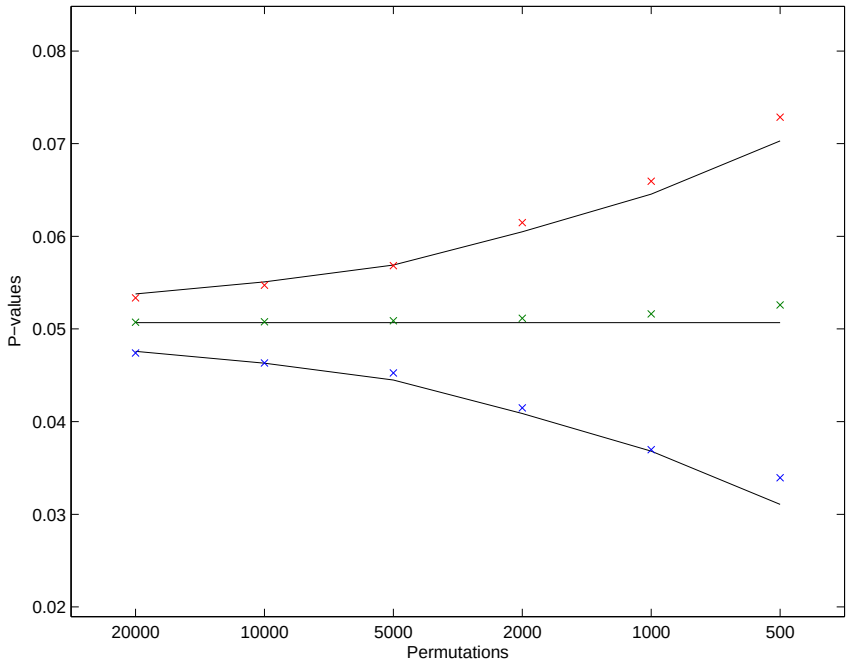


Figure 2.11: P-value precision as a function of number of permutations, with a parametric p-value of 0.05. See figure 2.10 for legend.

indicator variables (e.g. ones) in level l can be permuted in $n_l!$ ways without altering the statistic, which is the origin of equation 2.2.

In general linear models with nuisance-covariates, every permutation of the data that alters the pairing of points with distinct rows of the design matrix will typically result in different statistics. As a simple example, an ANCOVA design which adds a single continuous nuisance-covariate with n distinct values to a one-way ANOVA can result in all $n!$ permutations giving different statistics. The question therefore arises (for which credit is due to Thomas Nichols), whether permutations which merely alter the relationship with the confounds are useful, or whether sampling the random permutations only from the set that change the interest covariates produces results which are better in some way (for example, higher power, or a tighter spread of p-values or α).

A related issue is the importance of the permutations themselves being unique, or in other words, being sampled without replacement from the complete set. This is relevant to parallel implementation of permutation testing software, since it is inefficient for multiple parallel nodes to swap data describing the set of random permutations they have chosen, in order to ensure that all nodes' pooled permutations are unique. Alternatively, consider the use of designs with imaging covariates [69]. If either the nuisance or interest vary over the voxels, then so too will the set of useful permutations. Maximum-distribution based FWE control using methods which involve spatial-extent, such as cluster-volume, cluster-mass [29], or cluster enhancement [30] clearly require the same set of permutations to be used at each voxel. If the set of permutations is chosen for a 'typical' design matrix,[40] it is quite likely that this could result in non-unique permutations of the design (or of just the interest) for some voxels. The effect of this is likely to be similar to having redundant permutations in a simple model. We therefore identify three classes of permutation which may be of interest, summarised in table 2.14. Below, we explore the accuracy and power that result from these different permutation classes in two sets of Monte Carlo simulations.

Class	Description
$\mathcal{P}_1$	$\{S : SZ \neq Z\}$, sampled without replacement
$\mathcal{P}_2$	$\{S : SR_0Z \neq R_0Z\}$, sampled without replacement
$\mathcal{P}_3$	$\{S\}$, sampled with replacement

Table 2.14: Permutation classes, from which N_p permutations are sampled.

$\mathcal{P}_1$ versus $\mathcal{P}_2$ in ANCOVA

We first investigate the effect of including permutations from $\mathcal{P}_2$ which are redundant in terms of $\mathcal{P}_1$, within the context of a simple ANCOVA example. Consider a balanced two-group design, with $n_1 = n_2 = 8$ giving ${}_{16}C_8 = 12,870$ distinct permutations of the categorical interest covariate. We then add to this a single continuous nuisance-covariate, after which we can sample from the permutation classes. The chosen number of permutations will affect the expected proportion of redundancy, so we consider two values, 5000 and 10,000, which can respectively be considered small and large with respect to

[40] For example the average over all voxels, though this choice is outside of the present scope.

the number of available permutations in $\mathcal{P}_1$. It is plausible that the relationship between the interest- and nuisance-covariates will impact on the relative merits of the permutation classes, so we investigate three situations for the nuisance-covariate: 'Standard' — a uniform random vector; 'Correlated' — the same added to the interest covariate, to create a relatively high degree of correlation; 'Outlier' — as in (i), but with a pseudo-outlier added, as described earlier.

We wish to investigate both size and power, over a range of common *a priori* significance levels: $\alpha_0 = \{0.01, 0.05, 0.1\}$. For power, we set the value of b_1 to 2 (chosen to give a range of powers from approximately 20% for correlated nuisance at $\alpha_0 = 0.01$ to just below 100% for uncorrelated nuisance at $\alpha_0 = 0.1$). For size, b_1 is zero, and b_0 is 1 for both size and power.

In each of these 12 scenarios, we simulated 100 different nuisance-covariates, each resulting in different sets of N_p permutation samples. Each of these were then evaluated over 100 different samples of noise (data-sets). The choice of permutation strategy (e.g. Freedman-Lane or Smith's method) is not of particular interest here, so we limit ourselves to Anderson & Robinson's Exact method [40] and its closest practical implementation: Freedman-Lane.

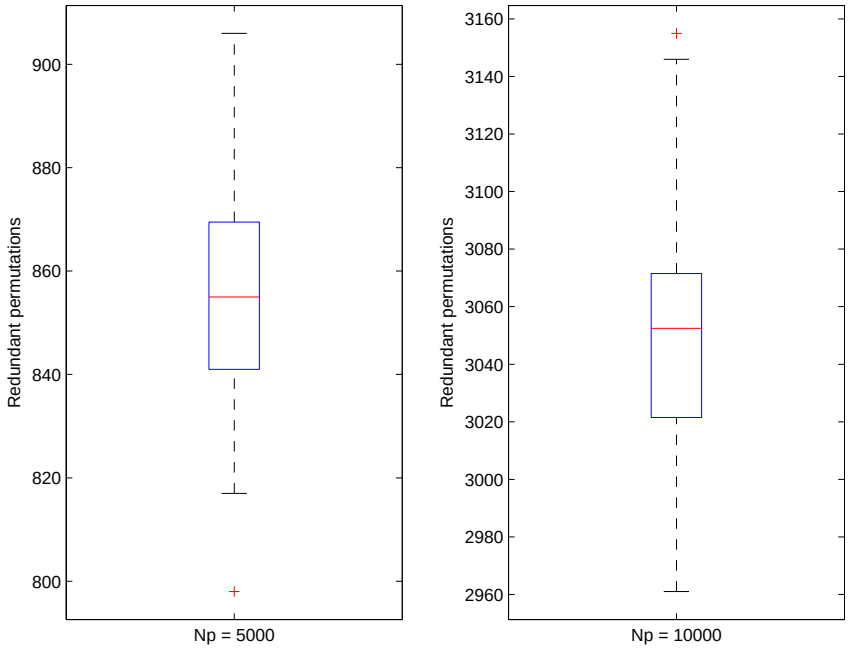


Figure 2.12: Numbers of permutations in $\mathcal{P}_2$ which are redundant under the terms of $\mathcal{P}_1$, over the 100 different simulated designs, for $N_p = 5000$ and $N_p = 10000$.

Figure 2.12 illustrates the level of '$\mathcal{P}_1$-redundancy' in $\mathcal{P}_2$, in terms of the number of samples from the latter which do not change the relationship between the data and the (original) interest covariate. Under this experimental set-up, there are 16! permutations in the complete set — such a large number that the sampling with replacement in $\mathcal{P}_3$ didn't actually result in any duplicate permutations.[41] Therefore, in this case, there is no

[41] In fact, continuing the simulation to sample a total of $1e7$ permutations from $\mathcal{P}_3$ only resulted in two duplicates.

difference between $\mathcal{P}_2$ and $\mathcal{P}_3$, so the latter will not be considered further with this design. A second experiment is described below, comparing $\mathcal{P}_3$ to $\mathcal{P}_1$ in a regression design with small n.

			Exact		Freedman-Lane	
Nuisance	$100\,\alpha_0$	N_p	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_1$	$\mathcal{P}_2$
Standard	1	5000	0.98	0.97	0.97	0.97
		10000	0.96	0.93	0.95	0.93
	5	5000	5.36	5.31	5.33	5.3
		10000	5.38	5.32	5.32	5.37
	10	5000	9.98	10.01	9.95	10.04
		10000	9.97	9.98	10.05	9.99
Correlated	1	5000	1.01	1.02	0.99	0.99
		10000	1.04	1.03	1.01	1.05
	5	5000	4.63	4.75	4.66	4.64
		10000	4.62	4.58	4.61	4.63
	10	5000	9.87	9.79	9.71	9.91
		10000	9.81	9.78	9.73	9.84
Outlier	1	5000	1.01	1	1.06	1.09
		10000	1	0.99	1.07	1.06
	5	5000	5.41	5.45*	5.36	5.38
		10000	5.4	5.39	5.31	5.42
	10	5000	10.33	10.4	10.35	10.39
		10000	10.38	10.29	10.41	10.39

Table 2.15: Effect of permutation class on accuracy, quantified by $100\,\alpha : \alpha' = \alpha_0$. Values outside the theoretical 95% confidence interval are starred. See table 2.14 for descriptions of $\mathcal{P}_i$.

We present results tables with a number of different performance metrics introduced in section 2.5.1, focussing on the smallest p-values or rejection rates, $\{0.01, 0.05, 0.1\}$, since these are of most interest. To quantify accuracy, we compare the observed size α to the expected size $\alpha' = \alpha_0$ given the true null hypothesis (table 2.15), highlighting any values outside the theoretical 95% confidence intervals based on 10,000 simulations.[42] Table 2.16 shows the bias in accuracy, averaging the error $\alpha - \alpha'$ over the ranges $0 \leq \alpha' \leq \alpha_0$, for each significance level α_0. To quantify the variability of the results from the different permutation classes, table 2.17 shows the root-mean-square of the above error over the same ranges. We also evaluate the expected uniformity of the p-values under the null hypothesis with the Kolmogorov-Smirnov statistic (table 2.18), again considering subsets of p-values in the ranges from 0 to the chosen α_0.

For evaluating power, we consider similar metrics to the first two above, but without the expected α', as this is unknown under the alternative hypothesis. Namely, we consider the mean (table 2.19) and standard deviation (table 2.20) of the observed values of α over the aforementioned ranges.

The main conclusion we draw from the results is that there appear to be few discernible patterns and no clear preferences. Both permutation classes are evidently valid, as demon-

[42] As given in table 2.13.

			Exact		Freedman-Lane	
Nuisance	$100\,\alpha_0$	N_p	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_1$	$\mathcal{P}_2$
Standard	1	5000	-0.0314	-0.0364	-0.029	-0.0556
		10000	-0.0345	-0.0277	-0.0365	-0.0494
	5	5000	0.109	0.1047	0.1076	0.09652
		10000	0.1058	0.1081	0.09662	0.09238
	10	5000	0.1394	0.1468	0.1374	0.1369
		10000	0.1393	0.1428	0.1412	0.1378
Correlated	1	5000	-0.0105	-0.0054	-0.0271	-0.0159
		10000	-0.0042	-0.0068	-0.0031	-0.002
	5	5000	-0.1515	-0.1538	-0.1542	-0.1462
		10000	-0.1503	-0.1453	-0.1559	-0.1417
	10	5000	-0.2682	-0.2634	-0.2709	-0.2615
		10000	-0.2652	-0.2822	-0.2653	-0.2617
Outlier	1	5000	-0.0341	-0.0144	-0.024	0.0091
		10000	-0.017	-0.0248	-0.0181	0.0166
	5	5000	0.1991	0.2061	0.1815	0.2318
		10000	0.1979	0.2056	0.1903	0.2362
	10	5000	0.2388	0.2597	0.2396	0.2643
		10000	0.2449	0.2489	0.2384	0.2836

Table 2.16: Effect of permutation class on accuracy, quantified by $100\,\text{mean}(\alpha - \alpha') : \alpha' \leq \alpha_0$, negative values are conservative. See table 2.14 for descriptions of $\mathcal{P}_i$.

			Exact		Freedman-Lane	
Nuisance	$100\,\alpha_0$	N_p	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_1$	$\mathcal{P}_2$
Standard	1	5000	0.04306	0.04512	0.03876	0.0593
		10000	0.04504	0.04117	0.04753	0.05452
	5	5000	0.1787	0.1841	0.1799	0.1735
		10000	0.1858	0.1804	0.1672	0.1716
	10	5000	0.1914	0.1965	0.1864	0.1879
		10000	0.1935	0.1882	0.188	0.1925
Correlated	1	5000	0.02373	0.02112	0.03599	0.02769
		10000	0.02383	0.02789	0.02398	0.02404
	5	5000	0.2023	0.1961	0.1923	0.1997
		10000	0.199	0.1952	0.2096	0.1922
	10	5000	0.3197	0.3034	0.3116	0.3111
		10000	0.3127	0.332	0.3086	0.3106
Outlier	1	5000	0.04498	0.02311	0.03636	0.0239
		10000	0.02905	0.03063	0.03211	0.0305
	5	5000	0.2519	0.2632	0.2376	0.2814
		10000	0.2518	0.2641	0.238	0.2865
	10	5000	0.2698	0.2949	0.2733	0.2939
		10000	0.2766	0.2843	0.2666	0.314

Table 2.17: Effect of permutation class on variability in test size, quantified by $100\sqrt{\text{mean}((\alpha - \alpha')^2)} : \alpha' \leq \alpha_0$, smaller is better. See table 2.14 for descriptions of $\mathcal{P}_i$.

			Exact		Freedman-Lane	
Nuisance	$100\,\alpha_0$	N_p	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_1$	$\mathcal{P}_2$
Standard	1	5000	0.09306	0.09979	0.0866	0.1082
		10000	0.09708	0.08473	0.1016	0.08742
	5	5000	0.04319	0.0405	0.0483	0.03412
		10000	0.04598	0.03636	0.04034	0.03994
	10	5000	0.03907	0.03657	0.04077	0.03392
		10000	0.04162	0.03901	0.03649	0.03952
Correlated	1	5000	0.07723	0.07412	0.09313	0.06869
		10000	0.08769	0.08359	0.07634	0.0781
	5	5000	0.03883	0.0349	0.03003	0.03274
		10000	0.03335	0.04287	0.03313	0.03431
	10	5000	0.05356	0.04036	0.03741	0.05241
		10000	0.04397	0.04912	0.04052	0.0482
Outlier	1	5000	0.1269	0.07	0.1091	0.0978
		10000	0.08	0.06525	0.1066	0.07434
	5	5000	0.0334	0.03366	0.02864	0.03368
		10000	0.02281	0.03684	0.03584	0.03279
	10	5000	0.02709	0.03235	0.02842	0.03097
		10000	0.02729	0.03319	0.02442	0.03235

Table 2.18: Effect of permutation class on p-value uniformity, quantified by ksstat (p vs uniform) : $p \leq \alpha_0$, smaller values indicate more uniform p-values. See table 2.14 for descriptions of $\mathcal{P}_i$.

			Exact		Freedman-Lane	
Nuisance	$100\,\alpha_0$	N_p	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_1$	$\mathcal{P}_2$
Standard	1	5000	59.98	60.1	59.92	60.25
		10000	60.78	60.66	60.62	60.99
	5	5000	82.22	82.32	82.26	82.32
		10000	82.45	82.44	82.41	82.54
	10	5000	88.87	88.95	88.91	88.92
		10000	89.01	89	88.98	89.05
Correlated	1	5000	10.85	11.01	10.98	11.55
		10000	11.03	11.08	11.06	11.71
	5	5000	28	28.17	28.13	28.58
		10000	28.13	28.13	28.14	28.66
	10	5000	39.06	39.2	39.17	39.47
		10000	39.17	39.16	39.2	39.53
Outlier	1	5000	60.39	60.48	60.32	61.91
		10000	61.18	60.97	61	62.63
	5	5000	83.01	83.04	82.96	83.47
		10000	83.2	83.15	83.15	83.65
	10	5000	89.49	89.49	89.47	89.73
		10000	89.59	89.55	89.57	89.83

Table 2.19: Effect of permutation class on power, quantified by $100\,\text{mean}(\alpha) : \alpha' \leq \alpha_0$, higher is better. See table 2.14 for descriptions of $\mathcal{P}_i$.

			Exact		Freedman-Lane	
Nuisance	$100\,\alpha_0$	N_p	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_1$	$\mathcal{P}_2$
Standard	1	5000	16.27	16.31	16.41	16.04
		10000	15.4	15.62	15.71	15.11
	5	5000	13.85	13.85	13.92	13.73
		10000	13.41	13.5	13.53	13.29
	10	5000	11.86	11.84	11.9	11.76
		10000	11.55	11.6	11.63	11.46
Correlated	1	5000	4.925	4.983	4.897	4.995
		10000	4.892	4.848	4.902	4.951
	5	5000	10.78	10.78	10.79	10.69
		10000	10.75	10.72	10.76	10.65
	10	5000	13.71	13.69	13.7	13.53
		10000	13.68	13.66	13.7	13.5
Outlier	1	5000	16.62	16.52	16.67	15.8
		10000	15.75	15.85	15.9	14.87
	5	5000	14.11	14.06	14.14	13.46
		10000	13.66	13.75	13.75	13.02
	10	5000	11.91	11.87	11.95	11.41
		10000	11.59	11.66	11.66	11.1

Table 2.20: Effect of permutation class on power variability, quantified by $100\,\mathrm{std}(\alpha)$: $\alpha' \leq \alpha_0$, smaller is better. See table 2.14 for descriptions of $\mathcal{P}_i$.

strated by the fact that all but one of the accuracies of the Exact and Freedman-Lane methods lie within the theoretical 95% confidence intervals.

Looking at the mean and root-mean-square errors in accuracy (tables 2.16 and 2.17), there seems to be no consistent winner. For example, considering the Exact method, $\mathcal{P}_1$ is preferred in 12 of the 18 cases in terms of bias, and in 10/18 cases in terms of variability. With Freedman-Lane, the equivalent results are 6 and 12/18 respectively. Under a null hypothesis of no preference between the permutation classes, a binomial distribution with 18 events of probability 0.5 gives [6,12] as an approximate 90% confidence interval, meaning none of these findings would be considered significant at a reasonable level. This binomial approximation is not strictly valid, however, because the rows of the table are not independent events — the results for different alpha are likely to be very strongly correlated, and the different N_p will probably have at least some correlation. Nevertheless, the conclusion that there is little evidence of a preference seems safe.

The other metrics seem to give similarly inconclusive results: The K-S statistic favours $\mathcal{P}_1$ 7/18 times for the Exact method, and 10/18 for Freedman-Lane. Interestingly, the average power of the Freedman-Lane method is slightly higher for $\mathcal{P}_2$ in all 18 cases, in contrast to our prior expectation that $\mathcal{P}_1$ should be superior. However, for the Exact method the number is only 10, which casts doubt on the significance of this finding. Power variability for Exact and Freedman-Lane prefers $\mathcal{P}_1$ 10 and 2 times out of 18, again giving weak evidence in favour of the class which includes permutations that change only the nuisance.

These results are limited, in that they only consider a single interest covariate, and a

single number n of data points. However, the design and n are fairly typical, and there is little reason to believe the distinction between $\mathcal{P}_1$ and $\mathcal{P}_2$ would be more pronounced with more complicated situations. At this stage, we must conclude that the experiment has produced no evidence to favour either permutation class.

$\mathcal{P}_1$ versus $\mathcal{P}_3$ in regression

We now consider the impact of having duplicate permutations in $\mathcal{P}_3$. A multiple regression setting is used with $n = 8$ giving $8! = 40,320$ possible permutations. The design consists of a single interest covariate and one nuisance-covariate in addition to a constant term. Both covariates are randomly sampled (from uniform $[0, 1]$ distributions) for 100 simulated designs, each evaluated over 100 data-sets, as above. Other details, for example the three classes of nuisance covariate, are the same as in the preceding experiment.

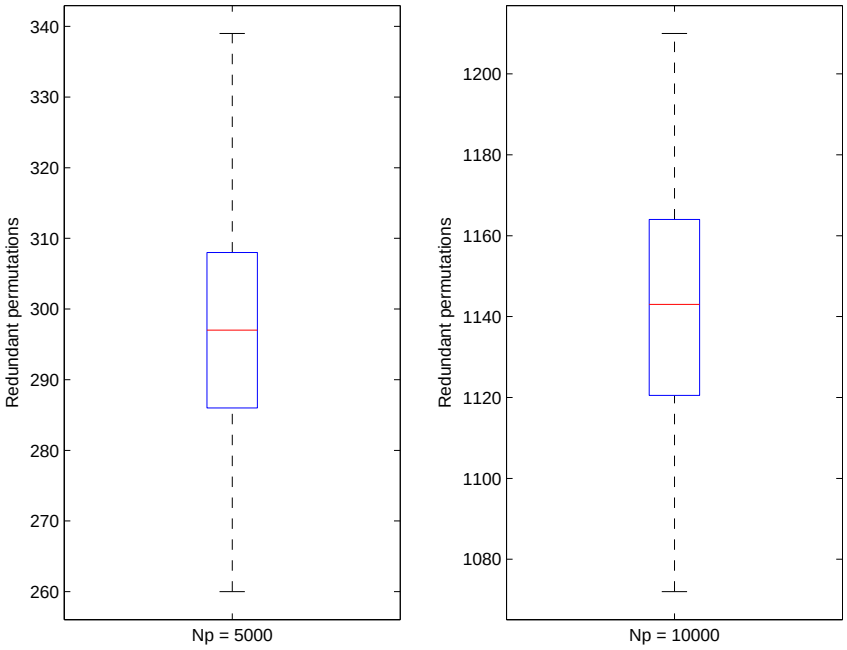


Figure 2.13: Numbers of permutations in $\mathcal{P}_3$ which are duplicates, over the 100 different simulated designs, for $N_p = 5000$ and $N_p = 10000$.

Figure 2.13 illustrates the number of duplicate permutations in the samples with replacement. We found all 100 random designs had distinct values for the original and orthogonalised interest covariate, meaning that the number of duplicates is also equal to the number of $\mathcal{P}_1$- and $\mathcal{P}_2$-redundant permutations sampled from $\mathcal{P}_3$.

The following tables present the results, in terms of size and power, as before.

In table 2.21 we observe ten values outside of the 95% confidence intervals; this is not very surprising, given the fact that 72 such comparisons have been performed without adjusting the confidence intervals for multiple comparisons. Additionally, the fact that the significant differences occur in four pairs, each including both $\mathcal{P}_1$ and $\mathcal{P}_3$, and that the two unpaired ones consist of one in each class means that there is no evidence that the class of permutation affects the validity.

			Exact		Freedman-Lane	
Nuisance	$100\,\alpha_0$	N_p	$\mathcal{P}_1$	$\mathcal{P}_3$	$\mathcal{P}_1$	$\mathcal{P}_3$
Standard	1	5000	1.08	1.03	1.05	1.1
		10000	1.04	1.08	1.04	1.06
	5	5000	4.81	4.83	4.8	4.89
		10000	4.79	4.79	4.82	4.78
	10	5000	9.68	9.73	9.74	9.84
		10000	9.69	9.71	9.78	9.78
Correlated	1	5000	1.2*	1.21*	1.16	1.1
		10000	1.22*	1.24*	1.13	1.08
	5	5000	4.74	4.77	4.82	4.78
		10000	4.73	4.82	4.8	4.81
	10	5000	9.56	9.49	9.58	9.59
		10000	9.59	9.55	9.65	9.62
Outlier	1	5000	1.14	1.19	1.2*	1.19
		10000	1.13	1.16	1.16	1.19
	5	5000	5.38	5.48*	5.51*	5.54*
		10000	5.39	5.34	5.48*	5.44*
	10	5000	9.95	9.98	10.14	10.12
		10000	10.01	9.99	10.11	10.03

Table 2.21: Effect of permutation class on accuracy, quantified by $100\,\alpha : \alpha' = \alpha_0$. Values outside the theoretical 95% confidence interval are starred. See table 2.14 for descriptions of $\mathcal{P}_i$.

			Exact		Freedman-Lane	
Nuisance	$100\,\alpha_0$	N_p	$\mathcal{P}_1$	$\mathcal{P}_3$	$\mathcal{P}_1$	$\mathcal{P}_3$
Standard	1	5000	0.0368	0.0011	0.0101	0.0112
		10000	0.0293	0.0282	0.0195	0.0284
	5	5000	-0.04906	-0.04254	-0.08452	-0.05786
		10000	-0.0385	-0.05	-0.06188	-0.05886
	10	5000	-0.2062	-0.2027	-0.1961	-0.1825
		10000	-0.2008	-0.218	-0.1749	-0.1878
Correlated	1	5000	0.0916	0.0733	0.0624	0.0568
		10000	0.1006	0.0805	0.0753	0.0611
	5	5000	0.00624	0.03082	0.00816	0.01548
		10000	0.02004	0.0252	0.01706	0.01004
	10	5000	-0.1834	-0.1602	-0.1847	-0.1784
		10000	-0.1818	-0.1679	-0.1852	-0.1746
Outlier	1	5000	0.0558	0.0521	0.0754	0.0809
		10000	0.0607	0.0604	0.0764	0.0625
	5	5000	0.1621	0.1761	0.2919	0.3168
		10000	0.1724	0.1425	0.2994	0.2732
	10	5000	0.1223	0.14	0.2418	0.2697
		10000	0.1389	0.1193	0.252	0.231

Table 2.22: Effect of permutation class on accuracy, quantified by $100\,\text{mean}(\alpha - \alpha') : \alpha' \leq \alpha_0$, negative values are conservative. See table 2.14 for descriptions of $\mathcal{P}_i$.

			Exact		Freedman-Lane	
Nuisance	$100\,\alpha_0$	N_p	$\mathcal{P}_1$	$\mathcal{P}_3$	$\mathcal{P}_1$	$\mathcal{P}_3$
Standard	1	5000	0.04519	0.02456	0.02777	0.0265
		10000	0.04026	0.04207	0.03437	0.03487
	5	5000	0.1065	0.1057	0.133	0.1212
		10000	0.1021	0.1179	0.121	0.1341
	10	5000	0.2716	0.2741	0.2406	0.2408
		10000	0.2707	0.291	0.2252	0.2492
Correlated	1	5000	0.113	0.09501	0.0749	0.07562
		10000	0.1212	0.106	0.08702	0.07748
	5	5000	0.1444	0.138	0.1033	0.09421
		10000	0.1465	0.1345	0.1071	0.09526
	10	5000	0.2867	0.2742	0.2847	0.2819
		10000	0.2954	0.277	0.2919	0.2704
Outlier	1	5000	0.07485	0.07421	0.09195	0.09647
		10000	0.07998	0.07979	0.08905	0.07853
	5	5000	0.1871	0.2175	0.3203	0.3507
		10000	0.1999	0.1665	0.3291	0.3031
	10	5000	0.1913	0.2191	0.2862	0.3096
		10000	0.1973	0.1786	0.2932	0.2729

Table 2.23: Effect of permutation class on variability in test size, quantified by $100\sqrt{\text{mean}((\alpha-\alpha')^2)} : \alpha' \leq \alpha_0$, smaller is better. See table 2.14 for descriptions of $\mathcal{P}_i$.

			Exact		Freedman-Lane	
Nuisance	$100\,\alpha_0$	N_p	$\mathcal{P}_1$	$\mathcal{P}_3$	$\mathcal{P}_1$	$\mathcal{P}_3$
Standard	1	5000	0.05593	0.08835	0.08	0.1073
		10000	0.06269	0.09444	0.05231	0.05472
	5	5000	0.04057	0.04282	0.03424	0.03831
		10000	0.04179	0.0454	0.03997	0.05007
	10	5000	0.02586	0.03727	0.02788	0.03765
		10000	0.02664	0.03811	0.02807	0.03777
Correlated	1	5000	0.07	0.07934	0.08103	0.06182
		10000	0.08	0.1	0.04646	0.08556
	5	5000	0.07335	0.06569	0.05	0.052
		10000	0.07484	0.05847	0.04683	0.05083
	10	5000	0.035	0.03363	0.02537	0.02412
		10000	0.03505	0.03087	0.02838	0.02508
Outlier	1	5000	0.08175	0.08168	0.08	0.07092
		10000	0.06761	0.06552	0.05966	0.08092
	5	5000	0.04292	0.05626	0.04066	0.04664
		10000	0.03789	0.04464	0.04801	0.03959
	10	5000	0.04462	0.05112	0.04854	0.05496
		10000	0.04348	0.0426	0.04723	0.04827

Table 2.24: Effect of permutation class on p-value uniformity, quantified by ksstat (p vs uniform) : $p \leq \alpha_0$, smaller values indicate more uniform p-values. See table 2.14 for descriptions of $\mathcal{P}_i$.

			Exact		Freedman-Lane	
Nuisance	$100\,\alpha_0$	N_p	$\mathcal{P}_1$	$\mathcal{P}_3$	$\mathcal{P}_1$	$\mathcal{P}_3$
Standard	1	5000	2.99	2.991	2.961	3.013
		10000	3.035	3.021	2.999	3.025
	5	5000	11.53	11.53	11.55	11.58
		10000	11.56	11.56	11.61	11.57
	10	5000	19.3	19.25	19.36	19.38
		10000	19.33	19.29	19.41	19.38
Correlated	1	5000	1.772	1.776	1.749	1.795
		10000	1.786	1.797	1.785	1.783
	5	5000	7.081	7.125	7.303	7.301
		10000	7.135	7.166	7.337	7.338
	10	5000	12.55	12.58	12.75	12.73
		10000	12.62	12.63	12.8	12.78
Outlier	1	5000	3.041	3.125	3.391	3.453
		10000	3.099	3.092	3.437	3.444
	5	5000	11.4	11.48	11.9	11.98
		10000	11.46	11.44	11.95	11.93
	10	5000	18.94	18.99	19.4	19.46
		10000	19	18.95	19.45	19.42

Table 2.25: Effect of permutation class on power, quantified by $100\,\text{mean}(\alpha) : \alpha' \leq \alpha_0$, higher is better. See table 2.14 for descriptions of $\mathcal{P}_i$.

			Exact		Freedman-Lane	
Nuisance	$100\,\alpha_0$	N_p	$\mathcal{P}_1$	$\mathcal{P}_3$	$\mathcal{P}_1$	$\mathcal{P}_3$
Standard	1	5000	1.674	1.688	1.692	1.688
		10000	1.692	1.685	1.698	1.745
	5	5000	5.688	5.693	5.735	5.727
		10000	5.683	5.687	5.74	5.711
	10	5000	9.118	9.077	9.177	9.166
		10000	9.105	9.076	9.174	9.169
Correlated	1	5000	0.9691	0.9448	0.9538	1.003
		10000	0.9521	0.9495	0.9711	0.9891
	5	5000	3.658	3.687	3.786	3.79
		10000	3.686	3.689	3.798	3.818
	10	5000	6.364	6.357	6.384	6.367
		10000	6.382	6.361	6.395	6.375
Outlier	1	5000	1.684	1.715	1.753	1.813
		10000	1.701	1.711	1.771	1.791
	5	5000	5.599	5.626	5.686	5.719
		10000	5.611	5.581	5.697	5.668
	10	5000	8.864	8.836	8.844	8.832
		10000	8.857	8.83	8.844	8.834

Table 2.26: Effect of permutation class on power variability, quantified by $100\,\text{std}(\alpha) : \alpha' \leq \alpha_0$, smaller is better. See table 2.14 for descriptions of $\mathcal{P}_i$.

Measure	Exact	F-L
Accuracy	6	7
Size var.	6	8
Uniformity	11	13
Mean power	7	8
Power var.	8	8

Table 2.27: Counts of the number of times out of 18 that $\mathcal{P}_1$ was preferred to $\mathcal{P}_3$ for various performance metrics, using Exact or Freedman-Lane permutation methods.

Table 2.27 summarises the results of pair-wise comparisons, as were discussed in greater detail for the previous experiment. Here, the results are slightly more consistent (with the exception of p-value uniformity quantified by Kolmogorov-Smirnov statistic), though somewhat counterintuitive. The permutation class with duplicates is typically preferred in around 10–12 of the comparisons. It is arguable here that variability of size and power might be expected to be lower with fewer 'real' permutations, so these are perhaps not good performance metrics. However, for $\mathcal{P}_3$ to give superior power more often than not is very surprising. Having said that, the balance is only slightly in favour, and the actual differences are very small, none of the percentage powers differ by more than 0.1. The actual values of power here are much lower than in the previous experiment, but this shouldn't be a particular problem, since there is no obvious pattern of better performance at higher or lower powers in either of the experiments. Further investigation is clearly needed here, as there must logically be a point at which duplication of permutations weakens the test. However, it seems fair to conclude that even with up to a third of the permutations being duplicates, there is no strong evidence of reduced performance. Sampling permutations with replacement is a step towards the (balanced) bootstrap [1], where this could occur, along with the sampling of more general combinations of the data which are not permutations. It is possible that theory from the bootstrap literature could help to explain the better-than-expected performance of the permutation test with redundant permutations.

2.6 Conclusions

In broad agreement with Anderson et al.'s theory [40] and simulations [8], we found the Freedman-Lane method to be a good approximation to Anderson and Robinson's hypothetical exact test. We conclude that it has the best overall balance of size and power. Ter Braak's method seems unable to offer better power without compromising size, so it cannot be recommended on the basis of the simulation results presented here.

Compared to Freedman-Lane, Smith's method showed similar performance in terms of size and power, and similar correlations between its set of permuted statistics and those of the exact method. It therefore seems that other authors [8, 40] may have been too quick to dismiss SZ and related methods like Sm. Furthermore, Sm seems uniformly preferable to SZ. Since it presents no additional difficulties in terms of O'Gorman's adaptive test [47, 48] it would seem logical that Sm should supersede the more basic method in this

context (where FL is inapplicable). The only practical disadvantage with Sm is that voxel-wise nuisance-covariates are slightly more expensive to handle than with SZ.

The transformed-residual strategies, including two novel variants of Huh and Jhun's method, were found to have excellent control of size under the null hypothesis, but disappointingly low power. They might be preferable to FL or Sm in cases where the noise distribution or other factors are particularly challenging, if the slightly higher susceptibility of FL to false positives is deemed more important than the lower power of the reduced-space permutation methods. However, further simulations would ideally be carried out to demonstrate that the differences can be significant, since only relatively minor departures of FL's size have been found here. However, there are two situations in which the transformed-residual strategies have a definite advantage over FL, firstly, they are valid for O'Gorman's adaptive test [47, 48]. Secondly, their equivalence for the pivotal t and non-pivotal $c^T b$ statistics, means that they should be valid for imaging tests which aim to control FWE via the maximum of a non-pivotal statistic , whereas FL's equivalence to Kennedy's method for $c^T b$ arguably discredits is use in such circumstances. We hypothesise a third situation in which transformed-residual strategies might be superior: with large r_0 and $n \gtrapprox r_0 + 7$. Here, the reduced permutation space will have an adequate (5000+) number of permutations, while the traditional methods will have a particularly large number of permutations for technically inexchangeable residuals. Further simulation experiments are needed to test this prediction.

Regarding the different classes of permutation, we have been unable to show any major differences between permutations that alter only the nuisance (or the orthogonalised interest) and those that alter the original interest-covariate(s). It also appears that sampling permutations with replacement, does not lead to the fall in power or increase in power variability that one would expect, though further simulations are needed to explore the limits of this phenomenon.

2.6.1 Further work

Several suggestions for future research have been given in the body of this chapter; the main ones are briefly recapitulated here, along with some additional suggestions. In section 2.3.4 two alternative methods for FWE control with combining functions were mentioned; we intend to perform Monte Carlo simulations and real-data evaluations to compare the practical performance of these techniques, and perhaps also to investigate the relative merits of combining p-values or raw statistics in the analysis of structural MRI data.

It has been shown that the method of Freedman-Lane is equivalent to the invalid method of Kennedy when they are both based on the un-normalised $c^T b$ statistic; in contrast, the transformed-residual methods have been noted to be equivalent for t and $c^T b$ due to the anicillarity of the nuisance [50]. In the case of neuroimaging studies, there is interest in studying the un-normalised stastistics, since they are less confounded by spatial smoothing [70]. We hypothesise that the use of a transformed-residual strategy in combination with step-down control of FWE (section 2.3.1) could lead to a method of inferring significance directly on the contrast images which is more principled and more

powerful than existing approaches.

As discussed in section 2.5.1, only a relatively modest number of simulations could be performed while keeping computation time and ease of interpretation at reasonable levels. We attempted to cover a wide range of potentially important scenarios, but must nevertheless admit that others could be of interest. Perhaps the most serious omission is that the simulations reported here did not consider multivariate data with non-Gaussian error. This is a complex issue though, which warrants an extensive set of simulations in its own right. Furthermore, only bivariate data was considered. This is something of a special case for multivariate data, as can be noted from the fact that Rao's F 'approximation' is actually exact for $m = 2$ (see appendix A.4.4). Ideally, a large range of m would be investigated, with an attempt to characterise the interactions between m, n and the numbers of interest and nuisance-covariates. For example, it might be expected that the typically robust large DF_E situation could still be fragile as m approaches n in magnitude. However, our experience here suggests that difficulties of interpreting the large numbers of results would make such an investigation quite challenging, even if computational resources were available.

All the simulations considered here have effectively been for individual voxels, in the sense that the maximum-distribution method of FWE control has not been investigated. Given that the subset pivotality assumption should be satisfied for images of voxel-wise statistics, it is safe to assume that any method which is valid for individual voxels will remain valid in the FWE case. Furthermore, it seems reasonable to believe that the relative ranking of methods in terms of power should be approximately maintained in moving from voxel-wise to image-wise analysis, since voxels where the alternative hypothesis holds will result in a greater maximum for the original permutation compared to the other permutations, while the null voxels will contribute similar statistics in the original and permuted arrangements. The converse, however, is less clear: it might be the case that a permutation method which does not control type I error for individual voxels could still prove to be reasonable in the imaging setting, since the distribution of the maximum over permutations could still control FWE. This should be the subject of further Monte Carlo simulations, and perhaps further theoretical consideration.

In contrast to the rest of the thesis, this chapter has not considered longitudinal data. Serial imaging will lead to within-subject correlations that mean the data is not exchangeable over time. However, there are several methods that have the potential to deal with this. Firstly, the special case of paired tests has already been discussed in section 2.4.1. More generally, with designs including both repeated-measures factors and exchangeable between-subject factors, the use of exact restricted-permutation strategies and/or permutation of 'exchangeable units' may be employed [36, 37]. More simply, using the multivariate GLM, balanced repeated-measures data (or summary statistics such as slopes and curvatures) can be analysed directly with the methods discussed here. These, and other approaches, including attempts to model serial autocorrelation [71] or to work with decorrelated (e.g. wavelet-transformed) data [72] should be investigated for application to longitudinal structural MR imaging.

Bibliography

[1] B. Manly, *Randomization, Bootstrap And Monte Carlo Methods in Biology*, 3rd ed. Chapman & Hall/CRC, 2006. ˆ60, 61, 62, 65, 66, 72, 73, 76, 77, 94, 95, 96, 97, 124, 136

[2] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993. ˆ60

[3] P. Good, *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer series in statistics, 2000. ˆ60

[4] T. E. Nichols and A. P. Holmes, "Nonparametric permutation tests for functional neuroimaging: A primer with examples," *Human Brain Mapping*, vol. 15, no. 1, pp. 1–25, 2002. ˆ61, 62, 67, 70

[5] T. Nichols and S. Hayasaka, "Controlling the familywise error rate in functional neuroimaging: a comparative review." *Stat Methods Med Res*, vol. 12, no. 5, pp. 419–446, Oct. 2003. ˆ61, 65, 66, 67, 79

[6] P. Good, "Extensions of the concept of exchangeability and their applications," *Journal of Modern Applied Statistical Methods*, vol. 1, pp. 243–247, 2002. [Online]. Available: http://tbf.coe.wayne.edu/jmasm/vol1_no2.pdf ˆ62, 73

[7] D. Commenges, "Transformations which preserve exchangeability and application to permutation tests," *Journal of Nonparametric Statistics*, vol. 15, no. 2, pp. 171–185, 2003. ˆ62, 65, 89, 92, 93, 94, 95

[8] M. J. Anderson and P. Legendre, "An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model," *Journal of Statistical Computation and Simulation*, vol. 62, no. 3, pp. 271–303, 1999. ˆ64, 65, 75, 76, 77, 94, 95, 96, 97, 101, 103, 120, 136

[9] P. Kennedy and B. Cade, "Randomization tests for multiple regression," *Communications in Statistics-Simulation and Computation*, vol. 25, no. 4, pp. 923–936, 1996. ˆ65, 73, 77, 79, 94, 95, 96, 99, 112

[10] J. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference*. CRC Press, 2003. ˆ65

[11] E. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks, Revised*. Prentice Hall, 1998. ˆ65

[12] E. Edgington, *Randomization Tests*. CRC Press, 1995. ˆ65, 73

[13] D. Pantazis, T. E. Nichols, S. Baillet, and R. M. Leahy, "A comparison of random field theory and permutation methods for the statistical analysis of MEG data." *Neuroimage*, vol. 25, no. 2, pp. 383–394, Apr. 2005. ˆ66, 67, 71

[14] Y. Hochberg and A. Tamhane, *Multiple comparison procedures*. John Wiley & Sons, 1987. ˆ67

[15] A. P. Holmes, "Statistical issues in functional brain mapping," Ph.D. dissertation, University of Glasgow, 1994. [Online]. Available: http://www.fil.ion.ucl.ac.uk/spm/doc/theses/andrew/ ˆ68

[16] M. Belmonte and D. Yurgelun-Todd, "Permutation testing made practical for functional magnetic resonance image analysis," *IEEE Trans. Med. Imag.*, vol. 20, no. 3, pp. 243–248, Mar. 2001. ˆ68

[17] W. Chau, A. R. McIntosh, S. E. Robinson, M. Schulz, and C. Pantev, "Improving permutation test power for group analysis of spatially filtered MEG data." *Neuroimage*, vol. 23, no. 3, pp. 983–996, Nov. 2004. ˆ68

[18] R. Johnson and D. Wichern, *Applied multivariate statistical analysis*, 5th ed. Prentice Hall, Upper Saddle River, NJ, 2002. ˆ69

[19] B. Cade and J. Richards, "Permutation tests for least absolute deviation regression," *Biometrics*, vol. 52, no. 3, pp. 886–902, 1996. [Online]. Available: http://www.jstor.org/stable/2533050 ˆ69

[20] M. J. Brammer, E. T. Bullmore, A. Simmons, S. C. Williams, P. M. Grasby, R. J. Howard, P. W. Woodruff, and S. Rabe-Hesketh, "Generic brain activation mapping in functional magnetic resonance imaging: a nonparametric approach." *Magn Reson Imaging*, vol. 15, no. 7, pp. 763–770, 1997. ˆ69

[21] C. Rorden, L. Bonilha, and T. E. Nichols, "Rank-order versus mean based statistics for neuroimaging." *Neuroimage*, vol. 35, no. 4, pp. 1531–1537, May 2007. ˆ69

[22] A. P. Holmes, R. C. Blair, J. D. Watson, and I. Ford, "Nonparametric analysis of statistic images from functional mapping experiments." *J Cereb Blood Flow Metab*, vol. 16, no. 1, pp. 7–22, Jan. 1996. ˆ69, 70

[23] J. B. Poline and B. M. Mazoyer, "Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters." *J Cereb Blood Flow Metab*, vol. 13, no. 3, pp. 425–437, May 1993. ˆ70

[24] K. Friston, K. Worsley, R. Frackowiak, J. Mazziotta, and A. Evans, "Assessing the significance of focal activations using their spatial extent," *Human Brain Mapping*, vol. 1, no. 3, pp. 210–220, 1994. ˆ70

[25] K. J. Friston, A. Holmes, J. B. Poline, C. J. Price, and C. D. Frith, "Detecting activations in PET and fMRI: levels of inference and power." *Neuroimage*, vol. 4, no. 3 Pt 1, pp. 223–235, Dec. 1996. ˆ70

[26] S. Hayasaka, K. L. Phan, I. Liberzon, K. J. Worsley, and T. E. Nichols, "Nonstationary cluster-size inference with random field and permutation methods." *Neuroimage*, vol. 22, no. 2, pp. 676–687, Jun. 2004. ˆ70

[27] J. B. Poline, K. J. Worsley, A. C. Evans, and K. J. Friston, "Combining spatial extent and peak intensity to test for activations in functional imaging." *Neuroimage*, vol. 5, no. 2, pp. 83–96, Feb. 1997. ˆ70, 71

[28] S. Hayasaka and T. E. Nichols, "Combining voxel intensity and cluster extent with permutation test framework." *Neuroimage*, vol. 23, no. 1, pp. 54–63, Sep. 2004. ˆ70, 71

[29] E. Bullmore, J. Suckling, S. Overmeyer, S. Rabe-Hesketh, E. Taylor, and M. Brammer, "Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain," *IEEE Trans. Med. Imag.*, vol. 18, no. 1, pp. 32–42, Jan. 1999. ˆ70, 71, 126

[30] S. M. Smith and T. E. Nichols, "Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference." *Neuroimage*, Apr. 2008. ˆ70, 126

[31] F. Pesarin, *Multivariate Permutation Tests: With Applications in Biostatistics*. J. Wiley, 2001. ˆ71

[32] N. A. Lazar, B. Luna, J. A. Sweeney, and W. F. Eddy, "Combining brains: a survey of methods for statistical pooling of information." *Neuroimage*, vol. 16, no. 2, pp. 538–550, Jun. 2002. ˆ71

[33] T. B. Terriberry, S. C. Joshi, and G. Gerig, "Hypothesis testing with nonlinear shape models." in *Inf. Process. Med. Imag.*, vol. 19, 2005, pp. 15–26. [Online]. Available: http://www.springerlink.com/content/e0d1jn4v28mc9qy9/ ˆ71, 72

[34] S. Hayasaka, A.-T. Du, A. Duarte, J. Kornak, G.-H. Jahng, M. W. Weiner, and N. Schuff, "A non-parametric approach for co-analysis of multi-modal brain imaging data: application to Alzheimer's disease." *Neuroimage*, vol. 30, no. 3, pp. 768–779, Apr. 2006. ˆ72

[35] R. Fisher, *The Design of Experiments*. Oliver & Boyd, Edinburgh, 1935. ˆ72

[36] M. Anderson and C. ter Braak, "Permutation tests for multi-factorial analysis of variance," *Journal of Statistical Computation and Simulation*, vol. 73, no. 2, pp. 85–113, 2003. [Online]. Available: http://www.informaworld.com/index/R15V17D6L0B8YKD8.pdf ˆ73, 75, 138

[37] J. Suckling and E. Bullmore, "Permutation tests for factorially designed neuroimaging experiments." *Hum Brain Mapp*, vol. 22, no. 3, pp. 193–205, Jul. 2004. ˆ73, 75, 95, 138

[38] H. Oja, "On permutation tests in multiple regression and analysis of covariance problems," *Australian & New Zealand Journal of Statistics*, vol. 29, no. 1, pp. 91–100, 1987. ˆ73, 95

[39] P. Kennedy, "Randomization tests in econometrics," *Journal of Business and Economic Statistics*, vol. 13, no. 1, pp. 85–94, 1995. [Online]. Available: http://www.jstor.org/stable/1392523 ˆ73, 76, 77, 79, 94, 95

[40] M. Anderson and J. Robinson, "Permutation tests for linear models," *Australian & New Zealand Journal of Statistics*, vol. 43, no. 1, pp. 75–88, 2001. ˆ74, 75, 76, 77, 80, 94, 95, 96, 101, 114, 118, 119, 124, 127, 136

[41] D. Freedman and D. Lane, "A nonstochastic interpretation of reported significance levels," *Journal of Business and Economic Statistics*, vol. 1, no. 4, pp. 292–98, 1983. [Online]. Available: http://www.jstor.org/stable/1391660 ˆ75, 95

[42] A. W. Still and A. P. White, "The approximate randomization test as an alternative to the F test in analysis of variance," *British Journal of Mathematical and Statistical Psychology*, vol. 34, no. 2, pp. 243–252, 1981. ˆ75, 94

[43] B. Jung, M. Jhun, and S. Song, "A new random permutation test in ANOVA models," *Statistical Papers*, vol. 48, no. 1, pp. 47–62, 2007. ˆ75, 83, 94, 95

[44] W. Welch, "Construction of permutation tests," *J. Amer. Statist. Assoc*, vol. 85, pp. 693–698, 1990. [Online]. Available: http://www.jstor.org/stable/2290004 ˆ75, 76

[45] C. ter Braak, "Permutation versus bootstrap significance tests in multiple regression and ANOVA," in *Bootstrapping and Related Techniques*. Springer, 1992, pp. 79–86. [Online]. Available: http://tinyurl.com/9yfbp3 ˆ76, 95

[46] B. Levin and H. Robbins, "Urn models for regression analysis, with applications to employment discrimination studies," *Law & Contemp. Probs.*, vol. 46, p. 247, 1983. [Online]. Available: http://www.jstor.org/stable/1191601 ˆ79, 95

[47] T. O'Gorman, "The performance of randomization tests that use permutations of independent variables," *Communications in Statistics-Simulation and Computation*, vol. 34, no. 4, pp. 895–908, 2005. ˆ80, 85, 94, 95, 136, 137

[48] T. W. O'Gorman, "An adaptive test of significance for a subset of regression coefficients." *Stat Med*, vol. 21, no. 22, pp. 3527–3542, Nov. 2002. ˆ80, 136, 137

[49] H. Theil, *Principles of Econometrics*. John Wiley, New York, 1971. ˆ80, 81, 86, 87, 89

[50] M. Huh and M. Jhun, "Random permuation testing in multiple linear regression," *Communications in Statistics-Theory and Methods*, vol. 30, no. 10, pp. 2023–2032, 2001. ˆ81, 82, 83, 86, 93, 94, 95, 103, 112, 113, 137

[51] S. S. Zocchi and B. F. J. Manly, "Generating different data sets for linear regression models with the same estimates," in *ICOTS*, vol. 7, 2006. [Online]. Available: http://www.stat.auckland.ac.nz/~iase/publications/17/C108.pdf ˆ82

[52] H. Theil, "The analysis of disturbances in regression analysis," *Journal of the American Statistical Association*, vol. 60, no. 312, pp. 1067–1079, 1965. ˆ87

[53] ——, "A simplification of the BLUS procedure for analyzing regression disturbances," *Journal of the American Statistical Association*, vol. 63, no. 321, pp. 242–251, 1968. [Online]. Available: http://www.jstor.org/stable/2283844 ˆ87

[54] W.-L. Luo and T. E. Nichols, "Diagnosis and exploration of massively univariate neuroimaging models." *Neuroimage*, vol. 19, no. 3, pp. 1014–1032, Jul. 2003. ˆ87

[55] S. Grossman and G. Styan, "Optimality properties of Theil's BLUS residuals," *Journal of the American Statistical Association*, vol. 67, no. 339, pp. 672–673, 1972. [Online]. Available: http://www.jstor.org/stable/2284464 ˆ87, 88

[56] G. C. Chow, "A note on the derivation of Theil's BLUS residuals," *Econometrica*, vol. 44, no. 3, pp. 609–610, 1976. ˆ88

[57] J. Magnus and A. Sinha, "On Theil's errors," *The Econometrics Journal*, vol. 8, no. 1, pp. 39–54, 2005. ˆ88, 89, 91, 92

[58] J. Schott, *Matrix analysis for statistics*. Wiley, New York, 1997. ˆ90

[59] C. Huang, "A test for multivariate normality of disturbance terms," *Southern Economic Journal*, vol. 38, no. 2, pp. 206–209, 1971. [Online]. Available: http://www.jstor.org/stable/1056831 ˆ90

[60] D. Naik, "Detection of outliers in the multivariate linear regression model," *Communications in Statistics-Theory and Methods*, vol. 18, no. 6, pp. 2225–2232, 1989. ˆ90

[61] F. Kianifard and W. Swallow, "A review of the development and application of recursive residuals in linear models," *Journal of the American Statistical Association*, vol. 91, no. 433, 1996. [Online]. Available: http://www.jstor.org/stable/2291419 ˆ90, 91

[62] J. Haslett and S. Haslett, "The three basic types of residuals for a linear model," *International Statistical Review*, vol. 75, no. 1, pp. 1–24, 2007. ˆ90

[63] H. Tobing and C. McGilchrist, "Recursive residuals for multivariate regression models," *Australian & New Zealand Journal of Statistics*, vol. 34, no. 2, pp. 217–232, 1992. ˆ91, 92

[64] H. Vinod, "Comments on "bootstrapping time series models"," *Econometric Reviews*, vol. 15, no. 2, pp. 183–190, 1996. ˆ93

[65] M. Grenier and C. Léger, "Bootstrapping regression models with BLUS residuals," *Canadian Journal of Statistics*, vol. 28, no. 1, pp. 31–44, 2000. [Online]. Available: http://www.jstor.org/stable/3315880 ˆ93

[66] A. E. Beaton, "Salvaging experiments: interpreting least squares in non-random samples," in *Computer Science and Statistics: Tenth Annual Symposium on the Interface*, D. Hogben and D. Fife, Eds. U.S. Department of Commerce, Washington, 1978, pp. 137–45. ˆ94

[67] E. Pitman, "Significance tests which may be applied to samples from any populations. II. the correlation coefficient test," *Journal of the Royal Statistical Society*, vol. B4, pp. 225–232, 1937. [Online]. Available: http://www.jstor.org/stable/2983647 ˆ95

[68] E. Edgington, *Statistical inference: the distribution-free approach.* McGraw-Hill, New York, 1969. ˆ123

[69] R. Casanova, R. Srikanth, A. Baer, P. J. Laurienti, J. H. Burdette, S. Hayasaka, L. Flowers, F. Wood, and J. A. Maldjian, "Biological parametric mapping: A statistical toolbox for multimodality brain image analysis." *Neuroimage*, vol. 34, no. 1, pp. 137–143, Jan. 2007. ˆ126

[70] M. Reimold, M. Slifstein, A. Heinz, W. Mueller-Schauenburg, and R. Bares, "Effect of spatial smoothing on t-maps: arguments for going back from t-maps to masked contrast images." *J Cereb Blood Flow Metab*, vol. 26, no. 6, pp. 751–759, Jun. 2006. ˆ137

[71] J. Locascio, P. Jennings, C. Moore, and S. Corkin, "Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging," *Human Brain Mapping*, vol. 5, pp. 168–193, 1997. [Online]. Available: http://www3.interscience.wiley.com/journal/56423/abstract ˆ138

[72] E. Bullmore, C. Long, J. Suckling, J. Fadili, G. Calvert, F. Zelaya, T. A. Carpenter, and M. Brammer, "Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains." *Hum Brain Mapp*, vol. 12, no. 2, pp. 61–78, Feb. 2001. ˆ138

Chapter 3

Voxel-Based Morphometry

Three distinct strands of work are presented in this chapter, all of them related to the technique of voxel-based morphometry. After a brief general introduction, we discuss a particular methodological issue that arises in the use of VBM to study neurodegenerative diseases such as Alzheimer's, and we develop an effective solution to this problem.

The second section specifically addresses the serial nature of the MRI data-sets with which this thesis is primarily concerned, proposing and evaluating new longitudinally-tailored preprocessing methods for VBM. To avoid the challenge of interpreting differing VBM results in the absence of ground truth, we use simulated data from a finite-element method (FEM) phenomenological model of atrophy (developed in collaboration with Camara et al. [1]).

Finally, we discuss a number of aspects regarding accurate and reproducible reporting of VBM methodology, and provide a set of recommendations (which were jointly developed by the coauthors of Ridgway et al. [2][1]) which we hope will help to standardise and advance this increasingly important field.

3.1 Introduction

Voxel-based morphometry (VBM), originally published by Wright et al. [3] but popularised by the methodological work of Ashburner and Friston [4] and the SPM software package (`http://www.fil.ion.ucl.ac.uk/spm`) is arguably the most successful among a number of techniques known in general as computational anatomy (CA) [5, 6, 7]. CA concerns the study of form and structure through mathematical and computational models, originally, in the work of Miller, Grenander, and co-authors [5, 6, 8] it referred quite specifically to modelling anatomical variation by means of a template and its associated deformation to model the individuals in a population, but the term has become more general in recent years [9].

In essence, Voxel-Based Morphometry involves Statistical Parametric Mapping (SPM) of data derived from structural MRI of multiple subjects. The images analysed are ob-

[1] In particular, Henley and Rohrer wrote parts of some of the rules, as well as discussing all of them, but the majority of the technical discussion as well as much of the final drafting of [2] were done by the present author.

tained through tissue segmentation, spatial normalisation, Jacobian modulation, and spatial smoothing [4]. Modulation refers to the procedure of multiplying the intensities in the spatially-normalised images by the Jacobian determinant of the transformation that maps from coordinates in the template space to those in the original image. There is occasionally some confusion about this procedure, so we briefly explain more carefully: in the approach used here, subject images are 'backward-mapped' into the template space so as to give a seamless result. This involves considering where each voxel in the warped result in template space should be taken from in the individual source. If a particular structure is larger in the subject than the template, the transformation models the expansion of the template, and hence has $|J| > 1$. Similarly, the process of spatial normalisation will geometrically shrink the subject's structure down to the size of the template. By multiplying the intensities in the warped result by the determinant, the original structure's volume is preserved. An alternative procedure is to forward-map the voxels from individual to template, superposing their contributions, which clearly also preserves volumes, though at the risk of creating seams in the warped result. This is the approach taken by Davatzikos et al. [10] in their RAVENS method, and in version 8 of SPM.[2]

Smoothing is used primarily to compensate for residual misregistration following spatial normalisation methods with limited flexibility [13]. It is also necessary for statistical parametric mapping in order to improve the normality of the data [14]. Lastly (and still potentially of importance even with very high-dimensional registration methods and non-parametric statistics), smoothing acts as a matched-filter [15], sensitising the analysis to a particular scale of effect.

3.2 Issues in masking for VBM of atrophy

3.2.1 Abstract

VBM performs voxel-wise statistical analysis of smoothed spatially normalised segmented MR Images. The analysis should include only voxels within a certain mask. We show that one of the most commonly used strategies for defining this mask runs a major risk of excluding from the analysis precisely those voxels where the subjects' brains were most vulnerable to atrophy. We investigate the issues related to mask construction, and recommend the use of an alternative strategy which greatly ameliorates this danger of false negatives.

3.2.2 Introduction

SPM performs mass-univariate statistical analysis using the general linear model at each voxel. More precisely, the calculations are performed at each voxel *within some mask*. There are several reasons why masking is necessary, mostly related to the multiple comparison problem. Family-wise error (FWE) correction using random field theory (RFT) is generally more powerful for smaller analysis regions (this is commented on further in the

[2] Though the DARTEL group-wise registration algorithm [11, 12] is often used for spatial normalisation in SPM8, which still uses backward-mapping.

discussion), and perhaps more importantly, masking is necessary for successful estimation of the smoothness of the residuals (John Ashburner, personal communication), which is a key part of the RFT correction procedure. If permutation methods [16] are employed for FWE correction, then the analysis region affects the computational complexity [17]. Correction of the false-discovery rate [18] also depends on masking, since non-brain voxels could otherwise skew the distribution of p-values on which it is based. Furthermore, masking can also partially alleviate a problem of implausible false positives occurring outside the brain due to the very low variance in voxels with consistently low smoothed tissue density — the extreme limit of the phenomenon described by Reimold et al. [19]. Finally, while not specifically considered here, multivariate machine learning, classification or decoding approaches [20, 21, 22] can also benefit from masking as an initial feature selection or dimensionality reduction step.

Having emphasised above that smaller masks generally lead to higher sensitivity and clarified interpretation, it is clearly important to recognise the obvious risk that overly restrictive masks will lead to false negatives, as potentially interesting voxels are excluded from the statistical analysis. In this section, we argue that there is a particular danger of false negatives arising in VBM studies of pathological brains when using the popular SPM software (`http://www.fil.ion.ucl.ac.uk/spm/`) with settings that are commonly used, and which appear reasonable a priori. We recommend the use of an alternative mask-generation strategy, which we show reduces this danger. In a three-part experiment, we will: use simulated data to investigate properties of preprocessing relevant to masking; explore the behaviour of standard and more novel methods of masking, considering variable patient group composition; and test the practical importance of our recommendation on a particular example of a real VBM study.

3.2.3 Methods

Masking strategies

The SPM software commonly used for VBM studies offers different alternatives to specify the mask for statistical analysis. If available, a precomputed mask can be explicitly requested, or the analysis mask can be automatically derived by excluding voxels in which any of the images have intensity values below a certain threshold. This threshold can be specified as an absolute value, constant for all the images, or as a relative fraction of each image's 'global' value. The global value can itself either be precomputed or can be automatically calculated as the mean of those voxel intensities which are above one eighth of the mean of all voxels. This arbitrary heuristic aims to determine an average that is not biased by the presence of potentially variable amounts of non-brain background in the field of view; it will be explored below.

In VBM studies where pronounced atrophy is expected, such as those of Alzheimer's disease (AD) [23] or semantic dementia (SD) [24], it is probable that some patients will have particularly low GM density in their most severely affected regions, but it seems undesirable to exclude such regions from the statistical analysis. Since this is likely to

occur with SPM's threshold masking, which effectively takes the intersection of all subjects' supra-threshold voxels, we recommend an alternative strategy for the creation of a mask (which can then be specified as an explicit mask in SPM). This strategy is based on the principle of replacing the criteria that all subjects should have voxel intensity above the threshold, with the relaxed requirement that some specified fraction of the subjects exhibit supra-threshold voxel values within the mask. In other words, voxels are included if there is a consensus among some percentage of the subjects that they are above threshold. Vemuri et al. [21] used this approach in their image classification work, with a consensus of 50% and a threshold of 0.1. SPM's method is a special case of this, where the consensus fraction is 100%. Another alternative masking strategy is to threshold the average of all subjects' segmentations; this might be expected to be similar to using a consensus of 50%, and this will also be explored. Software for VBM analysis has recently been released as part of the FMRIB Software Library [25], these scripts (`http://www.fmrib.ox.ac.uk/fsl/fslvbm/index.html`) implement another alternative procedure for their mask creation, which will be briefly evaluated.

Simulated images and optimality

In the first experiment, simulated images will be used from the BrainWeb project [26, 27], derived from real MRIs of normal healthy subjects. These images have known underlying tissue segmentation models, allowing quantitative evaluation of segmentation accuracy. By considering the simple Jaccard Similarity coefficient [28] between the discrete (maximum probability) model of GM and the estimated probabilistic segmentation after binarisation at a particular threshold, the optimal threshold may be found as a simple maximisation problem.[3] The optimal threshold will be investigated with relation to preprocessing, particularly smoothing, and in terms of its proportionality to the global or total signal. SPM's estimated global averages will be compared to a simple integrated total of the (probabilistic) voxel tissue volumes in litres.

AD cohort with varying composition

To provide a clearer characterisation of the effect of atrophy severity on mask construction, this experiment will consider different subsets from a typical VBM cohort of 19 patients with probable AD (M:F 9:10, mean age 68.8) and 19 healthy controls (M:F 8:11, mean age 68.3). All subjects were recruited from the Cognitive Disorders Clinic at the National Hospital for Neurology and Neurosurgery and gave written informed consent. They were assessed using standard diagnostic criteria including the Mini-Mental State Examination (MMSE) [29]. The study was approved by the local ethics committee. Imaging was performed on a 1.5 T GE Signa unit, using a spoiled fast gradient-recalled acquisition in the steady-state. 124 contiguous 1.5 mm-thick coronal slices with a 24 cm field of view and 256×256 matrix were acquired. The scan acquisition parameters were as follows: repetition time = 15 ms; echo time = 5.4 ms; flip angle = 15 °; inversion time = 650 ms.

[3] For example, using MATLAB's fminbnd to search for the best threshold between 0 and 1.

This experiment focusses on the robustness of the generated masks with respect to changes in the composition of the subject group, we will begin with only the 19 controls, before adding a single visually-severe AD patient and re-creating the masks, then finally the 18 remaining AD patients will be included, providing a typical balanced two-group comparison.

Practical importance on FTD example

Finally, the potential for overly restrictive masks to exclude potentially interesting findings in the most atrophied structures will be highlighted through presentation of the SPM results for a particular VBM study. A group of 14 fronto-temporal dementia (FTD) patients (M:F 7:7, mean age 63.5) with pronounced and focal temporal lobe atrophy, will be compared to a group of 22 approximately matched controls (M:F 10:12, mean age 65.8). All subjects were recruited from a specialist dementia clinic and gave written informed consent. They were assessed using standard diagnostic criteria. The study was approved by the local ethics committee.

Results will be presented following the masking strategy previously standard within our group, and with an example using the newly recommended consensus masking strategy, chosen based on visual evaluation of the suitability of various masks, prior to performing the statistical analysis.

3.2.4 Results and discussion

Simulated images

Figure 3.1 illustrates typical results for VBM preprocessing using SPM5's unified segmentation model [30]. The estimated segmentation is in close agreement with the simulation's underlying model,[4] but the inter-subject correspondence following spatial normalisation is only approximate. This imperfect overlap necessitates smoothing, but we can observe that even after smoothing there could be poor correspondence if the results were binarised with a relatively high threshold.

In figure 3.2, we explore the results from using the threshold which maximises the Jaccard similarity coefficient between the binarised probabilistic segmentation and the binary segmentation from the discrete BrainWeb model. It can be seen that while the original segmentation can be binarised successfully at a very high threshold, after spatial smoothing with an 8mm full-width at half-maximum (FWHM) Gaussian kernel, results are visually unacceptable with the theoretically optimal threshold. The threshold must clearly be lowered to include all desired voxels; fig. 3.2(f) shows the result from a much lower threshold, which, while often employed in VBM within our group, appears here to be far too generous. This apparent generosity should be contrasted with the findings shown later in figures 3.5 & 3.7.

[4] In fact, one of the most noticeable differences is that SPM's use of prior tissue probability maps has excluded some unrealistic dural 'GM' present in the simulation.

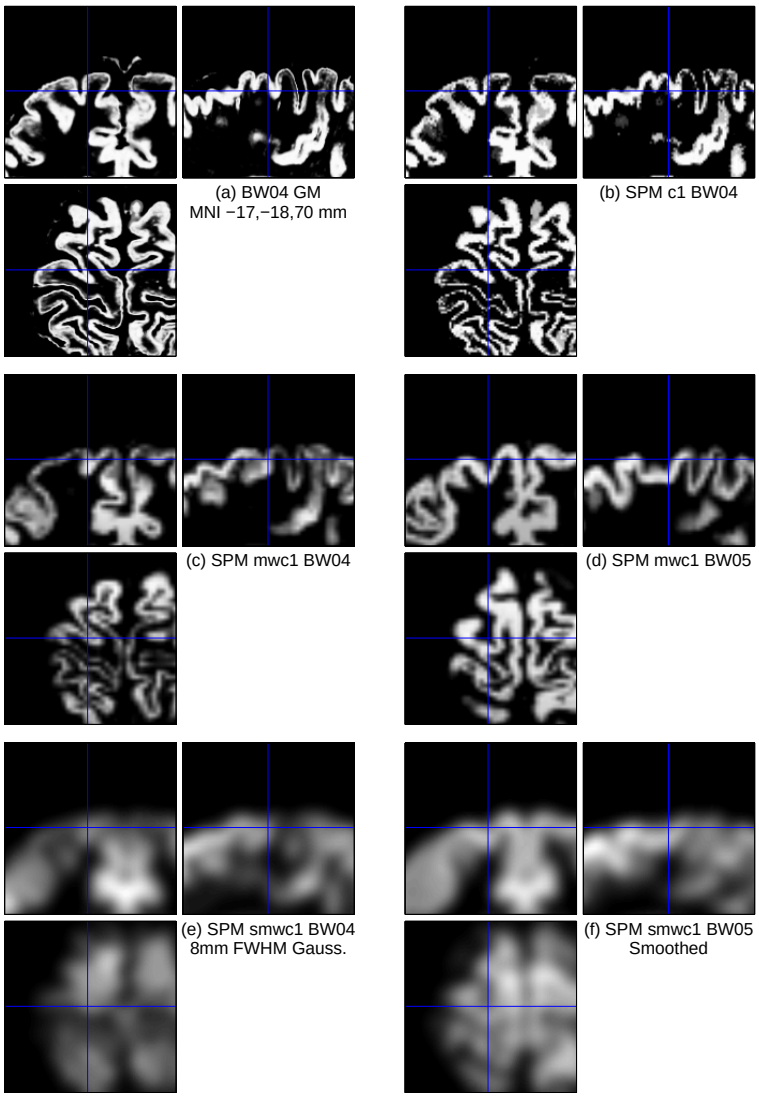


Figure 3.1: Illustration of the accuracy of tissue segmentation and inter-subject spatial normalisation, and the effects of smoothing. (a) and (b) compare the grey matter model used in the BrainWeb simulation to SPM's grey matter segmentation of the simulated T1 image. (c) and (d) show the anatomical correspondence between two different simulated subjects' results after spatial normalisation with a few thousand basis functions. (e) and (f) show the results following spatial smoothing.

To briefly investigate the use of relative thresholding, table 3.1 first compares the values of SPM's 'global' average to the totals from integrating over voxels, with three different sources of input data for four simulated subjects. It is clear that the total value is insensitive to the choice of these preprocessed source images, unlike the global value. Since the total also has the additional merits of being much simpler to interpret clinically, and of not using an arbitrary threshold (the 1/8 of the original mean), it seems preferable to use these totals as values for deriving proportional masking thresholds.[5] As a quick check of the suitability of these totals for relative threshold masking, table 3.2 presents

[5]While not evaluated here, it would also seem reasonable to prefer these more interpretable and stable values when adjusting for global volume in VBM, either through covariates or scaling-factors.

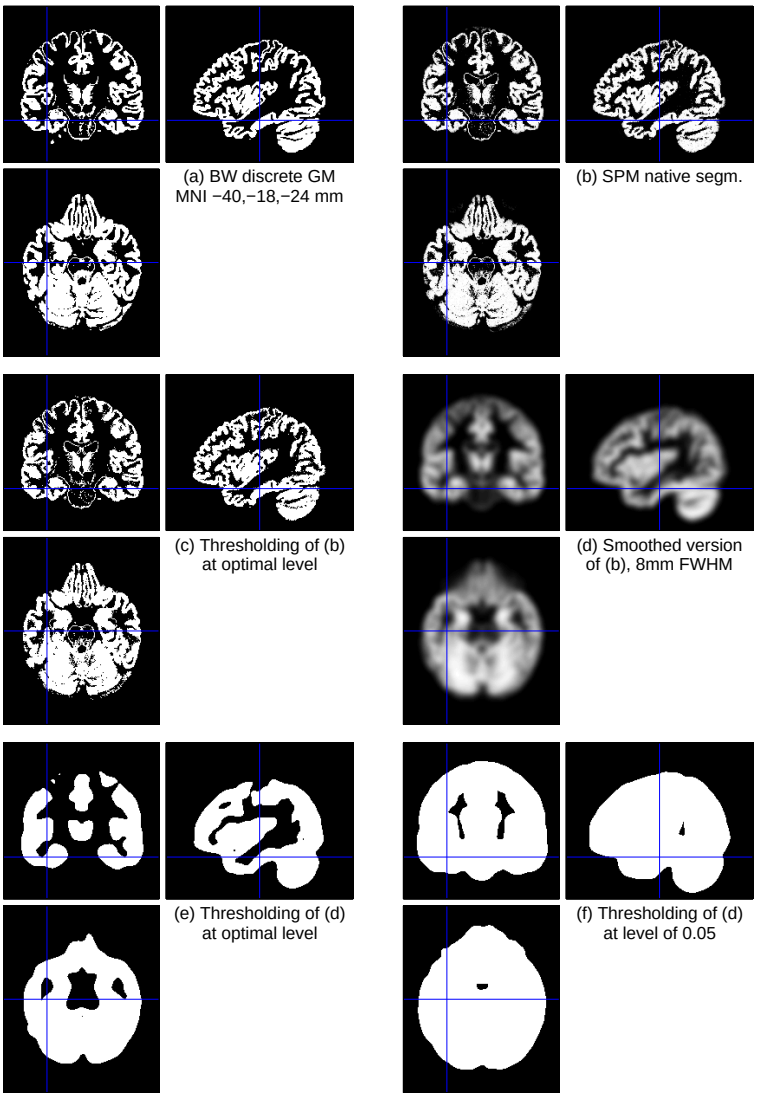


Figure 3.2: Optimal binarisation of probabilistic segmentations, and the interaction between smoothing and thresholding. (a) The binary GM label for BW, giving the voxels which have greater probability of being GM than any other tissue. (b) SPM's segmentation of the corresponding simulated T1 image. (c) SPM's segmentation thresholded at a level giving the optimal Jaccard similarity coefficient with the BW label. (d) Spatially smoothed SPM segmentation (8mm FWHM Gaussian). (e) and (f) comparison of thresholding of (d) at the optimal level and at a more typical absolute masking threshold.

the optimal absolute thresholds for four subjects, and the fractions of the global or total values necessary to achieve these thresholds. There is no apparent problem with using the totals, and limited evidence that they in fact have a more consistent relationship with the optimal threshold than the globals.

AD cohort

Continuing the comparison of SPM's globals with the integrated tissue totals from the previous experiment, figure 3.3 shows a strong correlation between these two measurements over all 38 subjects.

Table 3.1: Comparison of SPM's 'Global' averages with integrated totals (in litres) for four BrainWeb subjects, based on native GM segmentations, modulated warped segmentations without smoothing, and with 8mm FWHM smoothing.

	Global			Total		
Subject	Native	Mod. Warped	Smooth M. W.	Native	M. W.	Smooth M. W.
BW1	0.788	0.606	0.402	0.911	0.911	0.911
BW4	0.811	0.650	0.430	0.970	0.970	0.970
BW5	0.782	0.586	0.397	0.907	0.907	0.908
BW6	0.751	0.566	0.387	0.888	0.888	0.888

Table 3.2: Optimal thresholds, in terms of Jaccard similarity coefficient with BrainWeb model, as absolute values, relative fractions of SPM 'Globals' and of Totals in litres.

Subject	Opt. Abs. Thr.	Opt. Rel. G.	Opt. Rel. T
BW1	0.364	0.904	0.400
BW4	0.379	0.881	0.390
BW5	0.373	0.939	0.411
BW6	0.361	0.934	0.407

A visual example of the range of atrophy present in this subject-group is given in figure 3.4. On rough inspection, it might appear from the preprocessed images that the spatial normalisation and smoothing has adequately standardised even the most severe subject.

However, figure 3.5 presents a range of masks generated from four different strategies, on the three differently composed sub-groups. Table 3.3 gives the corresponding quantified mask volumes. It is clear from row (a) that the default SPM absolute thresholding strategy is very fragile with respect to the inclusion of atrophied patients. Adding the single severe individual results in a noticeably smaller mask, with particular reductions in the frontal and temporal lobes, and 100ml less total volume. The addition of the remaining 18 AD patients causes a further 120ml reduction in mask volume — corresponding to a loss of approximately 15,000 2mm isotropic voxels. Potentially interesting frontal cortex would not be analysed if such a mask was used. By lowering the consensus from SPM's 100% to 70%, the results become dramatically more robust to the inclusion of the patients. Row (b) of the figure shows only visually minor reductions in the mask; the table reveals that the volume loss through adding the severe case is just a tenth of that with the SPM strategy, though this rises to half when the remaining patients are added.

The use of relative thresholding should reduce the sensitivity to disease severity, since more severely atrophied patients will have lower global values and hence lower relative thresholds. However, example results (row c) using SPM's relative threshold masking (based on globals) still show a disturbing loss of cortical GM voxels from the mask with one patient, worsening with the additional patients. The overall loss when adding all patients to just the controls is 180ml (over 12% of the volume of the controls-only mask). Row (d) has the most visually appealing masks, derived from a 70% consensus and a threshold

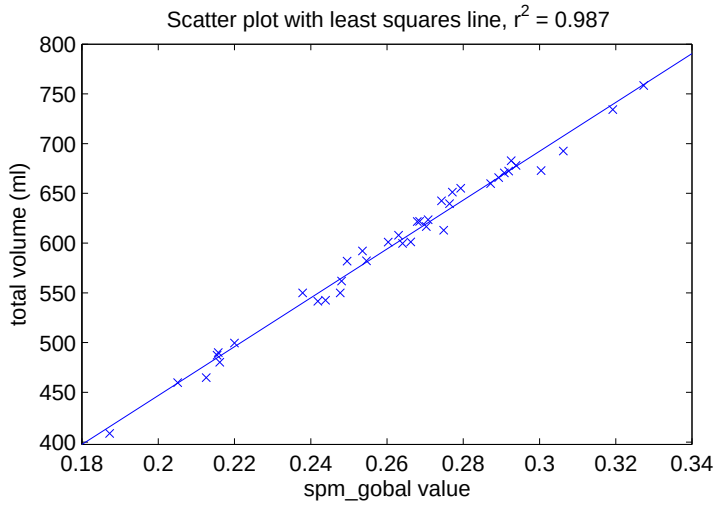


Figure 3.3: Comparison of total probabilistic GM volume in ml to SPM's 'global' values over 19 AD patients and 19 matched controls.

Table 3.3: Mask volumes (in litres) for the masks presented in Figure 3.5. See figure caption for row descriptions.

Method	19 Controls	Controls + 1 AD	all 38 subjects
(a)	1.79	1.69	1.57
(b)	1.97	1.96	1.90
(c)	1.47	1.40	1.29
(d)	1.63	1.62	1.59

relative to the integrated total volumes. The loss when adding all patients is now less than 2.5% of the original controls-only mask volume. It is self-evident in this experiment that the mask volume lost when adding a patient group to a control group could correspond to clinically-significant tissue loss in the actual control-patient comparison of interest. That this lost mask-volume can coincide with statistically-significant tissue differences is demonstrated in the next experiment.

Finally within this experiment, one potential problem with over-generous masks is demonstrated. In figure 3.6 some of the most significant voxels fall in regions where the majority of images do not have substantial chance of being genuine GM tissue. It is difficult to conclude confidently whether or not these are false positives, but the low variance and greater residual roughness present at the illustrated voxel certainly cast some doubt on the strength of the finding.

FTD example

The comparison of FTD patients with healthy controls reveals a pattern of tissue loss with focal left temporal lobe atrophy. Unthresholded SPM t-maps are shown in figure 3.7 (a) and (b); the two masks used for these analyses are overlaid in (c), where it is immediately obvious that the 100% consensus mask has excluded tissue in the temporal lobes,

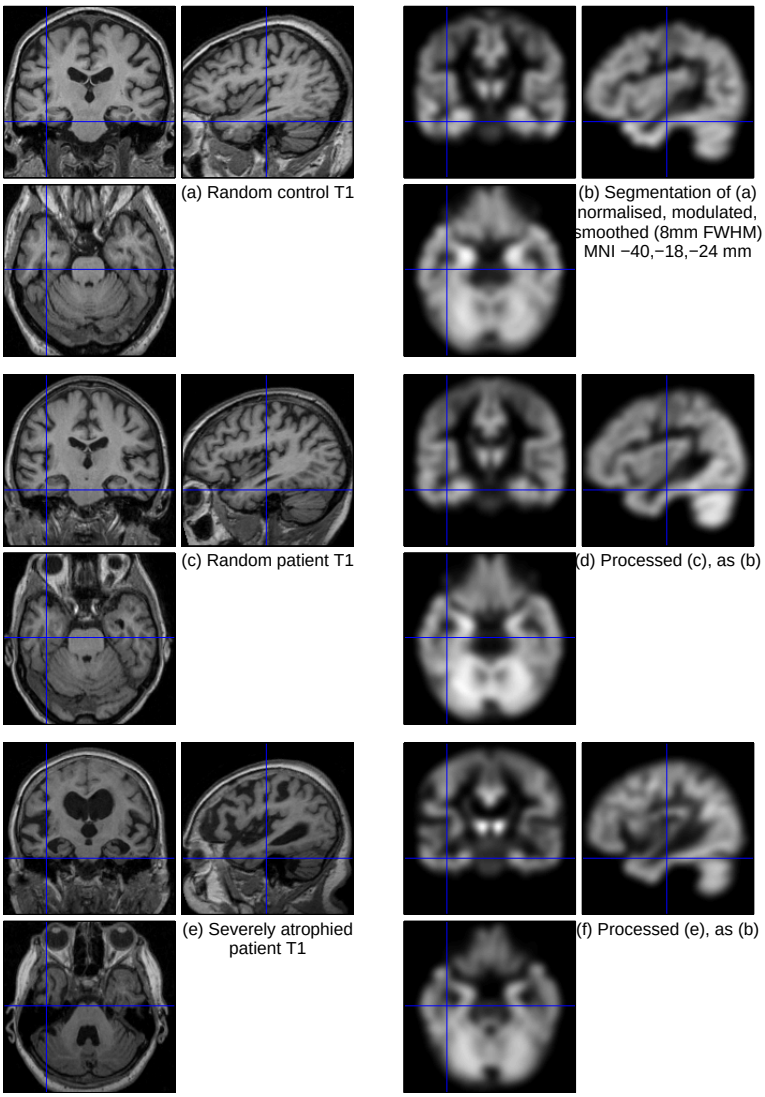


Figure 3.4: Example AD subjects. On the left are the standard clinical T1 images; on the right are their corresponding preprocessed segmentations. (a) and (b) are for a typical randomly selected healthy control from the group of 19. (c) and (d) are for one of the 19 patients, randomly chosen. (e) and (f) show the most severe patient, chosen in terms of visual assessment of overall tissue volume.

particularly on the left. The difference in volume of these two masks is over 300ml. Most importantly, (d) shows that some of the statistically-significant voxels (pFWE < 0.05) found when using the more reasonable mask will be ignored in the analysis using the standard 100% consensus mask. This lost significant volume amounts to 8.19ml, or over 1000 2mm isotropic voxels, in exactly the areas that these FTD brains are most atrophied.

Alternative masking strategies

Unlike the SPM strategies so far considered, which are derived from the smoothed (and optionally modulated) normalised segmentations, as used for the statistical analysis, FSL's

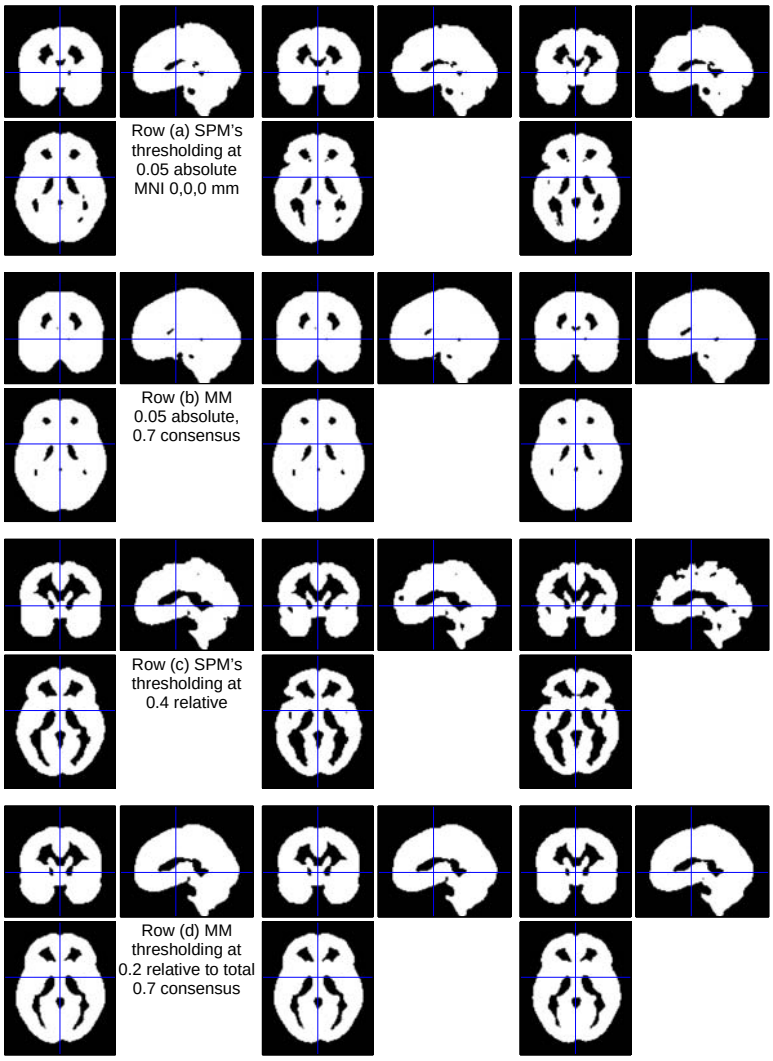


Figure 3.5: Masking results for varying method and patient group composition. The left column is for the group of 19 healthy controls; middle column, 20 images, including controls and the most severely atrophied patient; right column, entire collection of 19 controls and 19 patients. Rows (a) and (b) present masks based on absolute thresholding at a level of 0.05, first with SPM's default strategy, and in (b) with a "Majority Mask" (MM) requiring 70% of the images over threshold. Rows (c) and (d) investigate relative thresholds. (c) uses SPM's default strategy with thresholds of 0.4 times SPM's global values. (d) requires 70% of the images to exceed thresholds of 0.2 times the total value in litres.

VBM masking strategy[6] is based on unsmoothed and unmodulated segmentations (even when modulated data are analysed). FSL-VBM includes voxels in the mask if they meet both the following criteria: the maximum tissue probability over all subjects is at least 0.1; the minimum over the subjects is non-zero.

Examples of this strategy are illustrated for the AD data-set in figure 3.8. The most noticeable difference is that the use of unsmoothed segmentations leads to a much rougher mask. For correction of FDR or permutation-based FWE control, this roughness is unlikely

[6] `http://www.fmrib.ox.ac.uk/fsl/fslvbm/index.html`

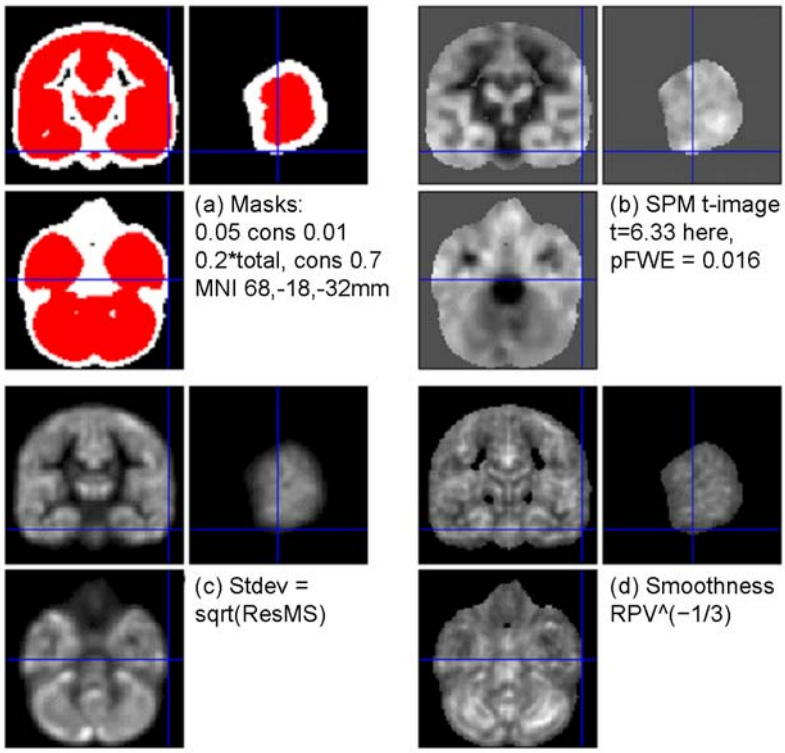


Figure 3.6: Masks and GLM results for the comparison of controls and AD patients. (a) shows the mask of Fig. 3.5(d, right column) overlaid on over-generous mask requiring only one of the images to exceed an absolute threshold of 0.05. (b-d) show the results from GLM estimation using this generous mask, in terms of t-values, standard deviation, and 'smoothness', respectively. The latter is derived from the 'resels per voxel' image, for easier visual interpretation.

to be a problem, but it may be detrimental for RFT-based correction of FWE. Worsley et al. [31] reported that expressions for RFT thresholding of statistics appeared to be most accurate for convex search regions, and they suggested that convoluted regions with high surface-area to volume ratios offer no advantage in power over smoother regions with larger volumes.

The lower panels of figure 3.8 show the results of applying FSL's mask inclusion criteria to the smoothed data which is actually analysed. In this case, smoothing leads to the presence of non-zero voxels as far away from the brain edges as the size of the support of the smoothing kernel used.[7] This effectively leaves only the second criterion in place; that the maximum over the segmentations be over 0.1. Now, we note that this criterion is simply a special case of the consensus masking strategy, where the consensus fraction is the reciprocal of the number of images, i.e. only one image (the maximum one for each voxel) need be above the threshold.

Finally, we consider deriving masks from the average of all subjects' smoothed normalised segmentations. This approach has been reported by Duchesne et al.[8] who binarised their average of 3mm FWHM smoothed unmodulated normalised segmentations

[7] In SPM5, the kernel is non-zero for ±6 standard deviations.

[8] Unpublished manuscript, available online: `http://www.bic.mni.mcgill.ca/users/duchesne/Proc/NI2004a.pdf`.

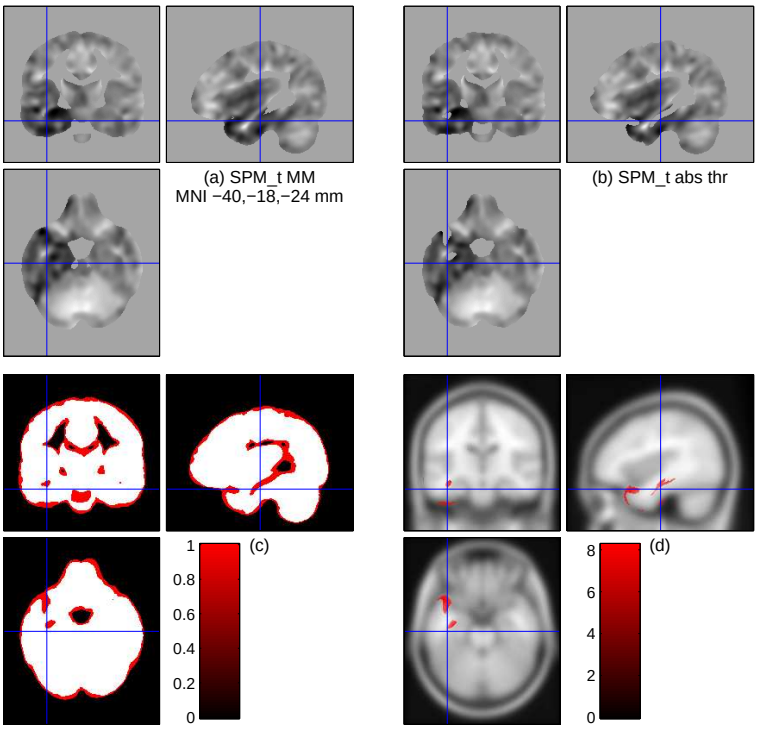


Figure 3.7: Masks and regions of significance (pFWE < 0.05) for the comparison of FTD subjects with controls. (a) and (b) show t-values for masking requiring either 70% (a) or 100% (b) of images to exceed a threshold of 0.05 (the latter corresponding to SPM's default strategy). (c) overlays the 100% mask on the 70% one. (d) overlaid on the group average segmentation is the region of significance present when using the 70% mask which is excluded from analysis with the default SPM masking strategy.

at a threshold of 0.3. Assuming that there is limited skew in the distribution of voxel intensities over subjects (SPM goes further in assuming normality), the arithmetic mean will approximately equal the median. Since the median by definition has 50% of the data beneath it, thresholding the average should be approximately equivalent to the special case of our proposed masking strategy with a consensus of 50% and the same threshold. In figure 3.9 we compare these two approaches on the AD data, showing almost identical results. As one would expect from the relatively low consensus fraction, there is very strong robustness to the addition of patients to the control group.

Based on the observation that the average image appears to have visually high probabilities over an intuitively reasonable region, it might be expected that a good threshold for the average would result in the binarised mask remaining highly correlated with the unthresholded original. One can determine an 'optimal' threshold such that this correlation is maximised. We compare such thresholds to higher and lower ones in figure 3.10, using the AD data-set. The volumes of the masks for the three subject groups are: 1.456, 1.456, and 1.444 litres — exhibiting a loss of below 1% with the addition of the AD patients. This provides a simple fully-automatic and objective technique for creating a mask, if the process of visual inspection and manual selection of multiple thresholds and/or consensus fractions is deemed too subjective.

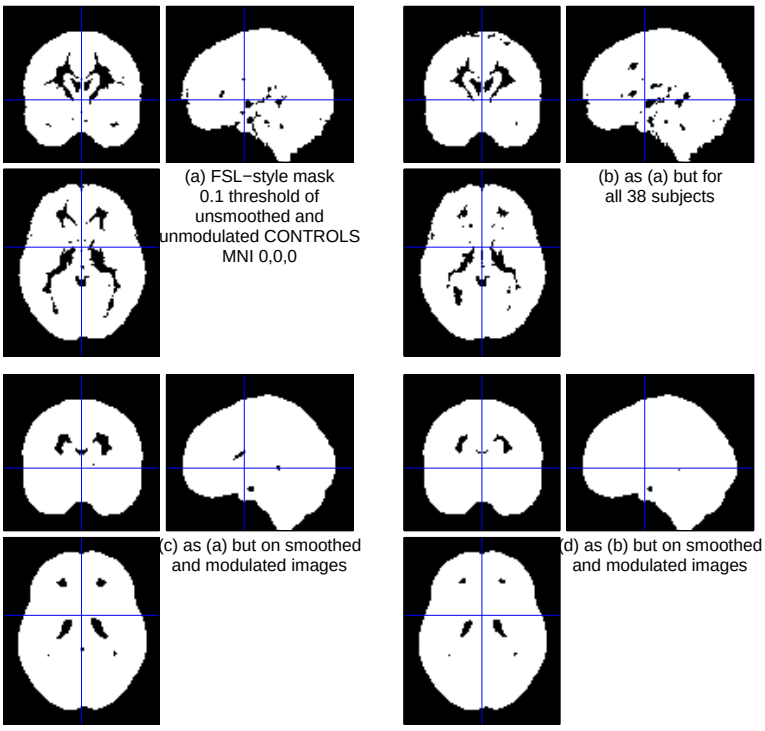


Figure 3.8: Comparison with the masking strategy used in FSL's VBM implementation. The top row derives masks from unsmoothed and unmodulated normalised segmentations, as in FSL; the bottom row uses smoothed modulated segmentations, as for the other SPM masking strategies discussed here. Left: for controls only; right: for all controls and AD subjects.

Further discussion

VBM studies aiming to localise small lesions or patterns of atrophy in finer scale structures require smaller smoothing kernels, due to the matched filter theorem [15]. The chance of losing interesting voxels from a mask created using absolute or relative thresholding with the standard 100% consensus is likely to be even greater with less smoothing. In a single subject with a severely atrophied small structure, greater amounts of smoothing would permit neighbouring tissue to bring the average value at the atrophied voxels above the threshold. However, it should also be noted that finer scale spatial normalisation [10, 11] may counterbalance this effect, as atrophied structures can be better warped to match those of the template/average, with the information about their atrophy being transferred to the deformation field.

Further work could involve extension of the method of automatic threshold selection, and/or selection of an optimal consensus fraction, perhaps using bootstrap methods or cross-validation. It may also be helpful to base masks upon the voxel-wise statistical results,[9] directly addressing the problem of false-positives in low-variance regions by excluding these voxels.

[9]SPM follows a related procedure for variance component estimation, only pooling over voxels which show main effects above a certain level of statistical significance [32].

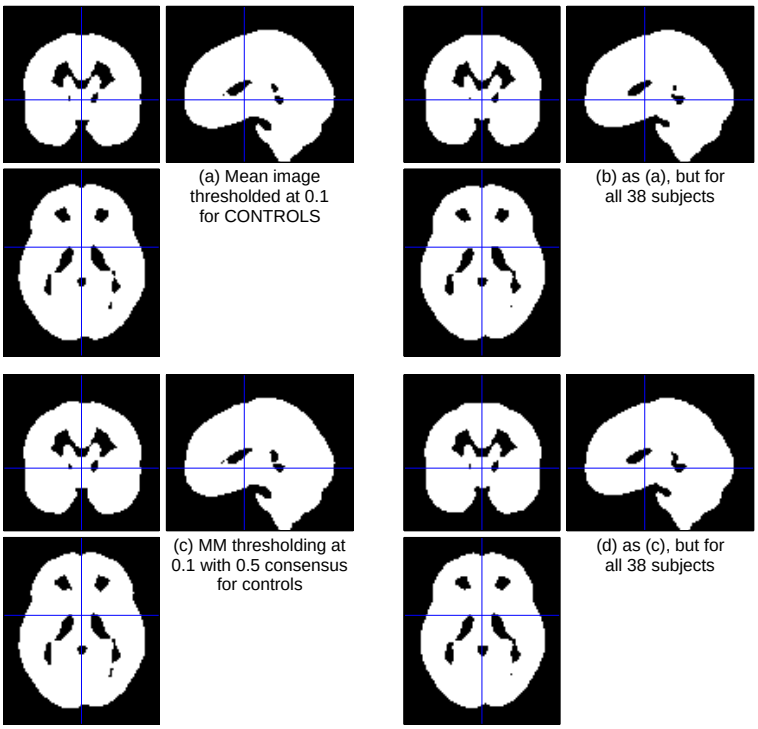


Figure 3.9: Top row: masks derived from thresholding (at 0.1) the group mean of the smoothed modulated normalised segmentations; bottom row similar masks using a 50% consensus of the unaveraged segmentations and the same threshold. Left: for controls only; right: for all controls and AD subjects.

3.2.5 Conclusions

The standard masking procedure in the SPM software risks missing findings in the most severely atrophied brain regions. It is important to note that the missed atrophy when using overly restrictive masks might not be readily apparent from consideration of the 'glass-brain' maximum intensity projection commonly presented in VBM results. It seems not to be standard practice for VBM papers to present the analysis region resulting from their choice of masking strategy. We would recommend careful checking of the mask, and would argue in favour of this occurring prior to the statistical analysis itself — a practice which is simplified by using the mask-creation strategy proposed here. We would additionally suggest that the masking procedure be reported clearly enough to be reproducible, as we have previously advocated [2]. Software is available to implement the consensus masking technique recommended here.[10]

[10] `http://www.cs.ucl.ac.uk/staff/gridgway/masking`

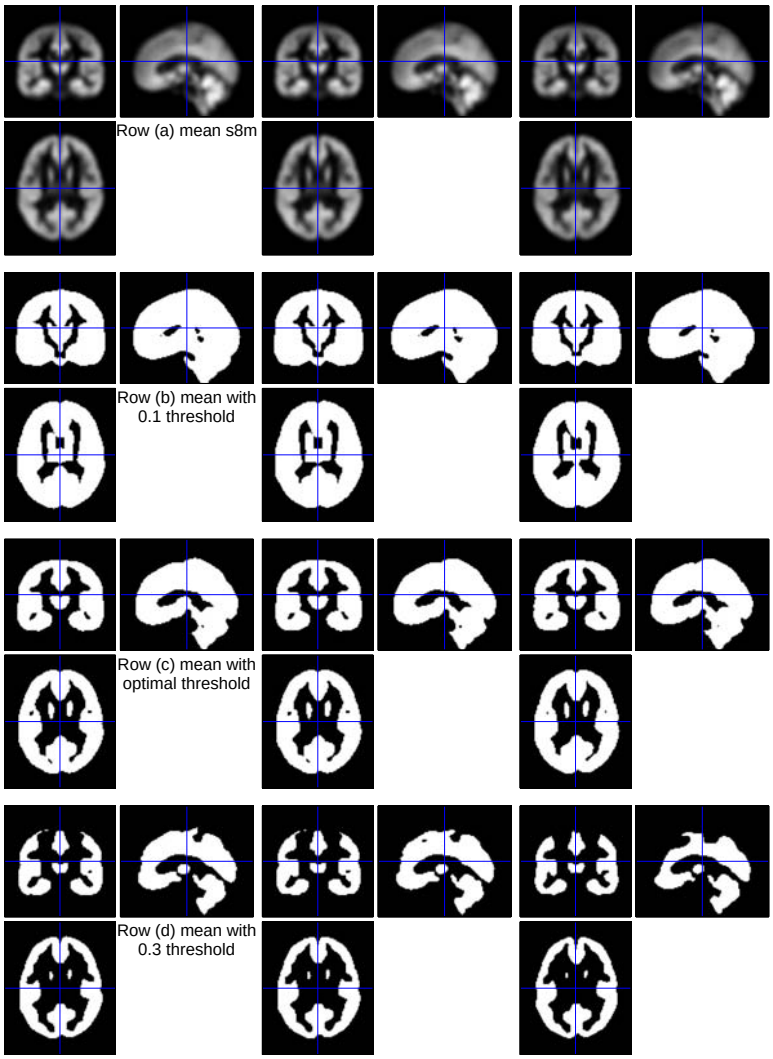


Figure 3.10: Further investigation of masks derived from the group average segmentation. Left column, for the control group; middle column, controls plus one severe AD patient; right column, all controls and AD subjects. Top row, the average segmentations themselves; row (b) the means thresholded at 0.1 (as in Fig. 3.9); row (c) the mean images thresholded at optimal levels of 0.203, 0.200, 0.189; bottom row, a higher than optimal threshold of 0.3.

3.3 Methods for longitudinal VBM

3.3.1 Abstract

The goal of this section is to evaluate Voxel-Based Morphometry and three longitudinally-tailored methods of VBM. To aid quantitative comparison of subtly different methods, images with simulated atrophy (and hence gold standard results) have been employed. Segmentation performance itself has also been directly evaluated within a small number of subjects derived from phantom brain images. The simulated atrophy images are produced by deforming original scans using a finite element method, guided to emulate Alzheimer-like changes based on the typical MR-observed disease-course and imaging-derived measurements of volume change. These simulated images provide quite realistic data with a known pattern of spatial atrophy, with which VBM's findings can be meaningfully compared. This is the first evaluation of VBM for which anatomically-plausible 'gold-standard' results are available.

The three longitudinal VBM methods have been implemented within the unified segmentation framework of SPM5; one of the techniques is a newly developed procedure, which shows promising potential for use with serially acquired structural imaging data.

3.3.2 Introduction

Longitudinal variants of VBM have been developed for application to cohorts with serial MR imaging [33, 34]. VBM necessitates preprocessing of the images, including spatial normalisation and tissue-segmentation. There are a number of options and adjustable parameters within the standard method, in addition to more dramatically different techniques such as longitudinal alternatives. In contrast to the ease of methodological tuning, there is great difficulty in evaluating the performance of VBM methods due to the lack of ground truth. To the best of our knowledge, no previously published VBM studies of realistically complex data have had gold-standard maps of the regions that should be detected.

We have developed finite element method (FEM) models which can structurally alter images, producing finely-controllable, clinically realistic changes [1]. Such simulated images have known underlying deformation fields and volume changes, which can form a gold standard for evaluating atrophy-measurement techniques.

Using a cohort of AD patients with MR images at baseline and one year later, we simulated new approximate year-on scans from the original baselines, guided by semi-automated measures of whole-brain, hippocampal, and ventricular volume changes [35].

The original baseline and simulated follow-up images then constitute a data-set with known FEM ground truth; we use this to derive a gold standard suitable for evaluating longitudinal VBM, and compare four such techniques, including our newly developed post-averaging procedure.

3.3.3 Methods

Voxel-Based Morphometry and Longitudinal VBM

The latest and most sophisticated version of segmentation for VBM (available in SPM5 and SPM8) unifies tissue-segmentation with spatial normalisation [30].

With serial data, statistical analysis can take advantage of reduced within-subject variability (e.g. using repeated-measures ANOVA on balanced data, or analysing longitudinal within-subject summary statistics). To capitalise on the longitudinal information, changes should also be made to the VBM preprocessing methods. In this work, we evaluate standard VBM against two longitudinal methods from the literature (which we have adapted to be compatible with the unified segmentation framework of SPM5) and our own newly developed SPM5 method.

For all four methods, SPM analysis was performed within an explicit mask, derived from the (smoothed) ground-truth grey-matter segmentation. This segmentation, the VBM subtraction images, and the equivalent gold-standard images (described below) were all smoothed with the same kernel — an 8mm full-width at half-maximum (FWHM) isotropic 3D Gaussian.

The balanced nature of the data (i.e. all subjects have equivalent time-points, at 0 and 12 months) means that within-subject summary statistics (for annualised change) are simply given by longitudinal subtraction images; these are entered into a one-sample t-test. A single-tailed contrast for atrophy (increase<0) and the 'reverse contrast' of tissue-gain (increase>0) were evaluated and thresholded with multiple comparison correction using random field theory (RFT) to control the family-wise error (FWE) at a 1% level.

Standard Here, 'Standard' VBM refers to simple application of unified preprocessing independently to each scan of each subject; only the statistics differ from the non-serial case. 'Standard' should not be contrasted here to 'optimised' VBM [36], which the unified segmentation model aims to supersede [30].

Tied-normalisation The preprocessing step of spatial normalisation should take advantage of the fact that multiple time-points for a single subject can be registered much more accurately than scans of different subjects, and that initial rigid alignment already reveals a great deal about within-subject change [37].

Using the non-unified model of SPM2, Gaser (in Draganski et al. [33]) developed a method with longitudinally tied spatial normalisation, in which repeat scans are transformed using the parameters determined for their corresponding baselines, then independently segmented.

Following the introduction of SPM5's unified framework, an extended generative model for unified longitudinal segmentation and normalisation should ideally be developed. As a simpler alternative, we have implemented an approach which applies the baseline normalisation parameters to the native-space baseline and follow-up grey matter images from separate unified segmentations.

Pre-averaged More advanced techniques can combine inter-subject spatial normalisation with precise intra-subject registration using High-Dimensional Warping (HDW). One such method (designed by Ashburner, and implemented in [34]), creates low-noise averaged images of HDW-registered longitudinal sets, before inter-subject spatial normalisation and segmentation in SPM2. (i.e. averaging is 'pre' segmentation.)

We have adapted this approach to the SPM5 framework, with unified segmentation and inter-subject normalisation following the intra-subject warping and averaging. The intra-subject volume changes from HDW must be taken into account to generate the follow-up data, which can be elegantly done by modulating the native-space segmented average-images with the HDW Jacobian fields before applying the predetermined inter-subject transformations. This avoids the interpolation-error due to the transformation of the Jacobians in [34].

Post-averaged We propose here a technique similar to pre-averaging, but novel, and better-suited to SPM5's unified segmentation. The new method should be superior for subjects with large longitudinal change that might not be fully recovered by HDW, as in this case, the pre-averaged images may be too blurred to segment well.

Each time-point is first segmented, and SPM5's bias-corrected version is saved; HDW transformations are then determined on the corrected images [11] and applied to their native-space segmentations. The warped segmentations are then averaged; i.e. averaging is 'post' segmentation of sharp original images. Each average segmentation is modulated with the HDW volume changes to create follow-up equivalents, and finally, each set of original and modulated segmentations is spatially normalised with the baseline parameters.

Finite Element Modelling of Atrophy

The atrophy simulation process is based on that described in [1]. It consists of four main steps: (1) Generation of a reference mesh; (2) Warping to a subject-specific mesh; (3) Deformation of the mesh using a FEM solver; (4) Application of the deformations to the baseline image of each subject, to produce a new simulated follow-up image. The reference mesh was built using the BrainWeb atlas labels of these structures [38] (`http://www.bic.mni.mcgill.ca/brainweb/`). The adaptation of the reference mesh to each subject was achieved with a mesh warping procedure guided by a fluid registration algorithm [39].

We used a cohort of 18 probable AD patients (7 female; ages from 55 to 86 years, mean 70) with baseline and 12-month follow-up MRI scans [35]. The FEM simulation was driven using values of the subjects' volume changes in the brain, hippocampi, and ventricles (from semi-automated segmentation-based measurements). Simulated mean (standard deviation) percentage volume increases were: brain, -2.43 (1.18); hippocampi, -4.74 (3.24); ventricles, 11.49 (5.35). Figure 3.11(a-c) shows a single-subject example of atrophy simulation; ventricular expansion, cortical thinning, and opening of CSF spaces can be observed.

[11] Bias corrected images were also used in the pre-averaged method, for fairness, though in practice this would require additional time-consuming unified segmentations, or more conventional bias correction with an external program.

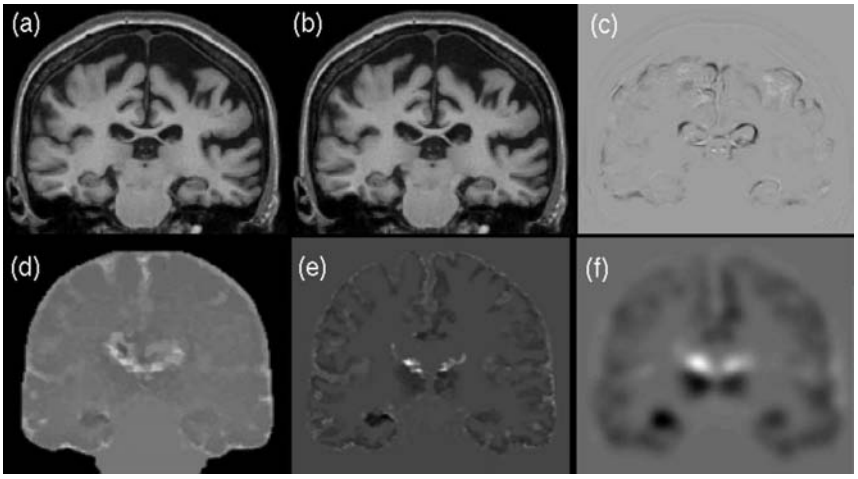


Figure 3.11: Example case of simulated atrophy: (a) Original baseline; (b) Simulated +1yr follow-up; (c) Subtraction image. The same subject's gold-standard volume changes in BrainWeb space: (d) Volume gain (VG=1yr/orig); (e) GM-increase = (GM*VG)-GM; and (f) Smoothed GM-increase, as entered into the analysis.

Generation of a Gold Standard Because the same mesh is warped from the BrainWeb template to each individual patient, there is a known correspondence between elements of the warped meshes for the different subjects; therefore the volume change of each element that results from the mesh-deformation can be mapped back to the common space. By converting the element-wise volume changes to a voxel-wise representation, an image of the ratio of follow-up volume to original volume is created.

These volume gain ratio images can be used to modulate the BrainWeb Grey Matter Segmentation, resulting in perfectly aligned effective follow-up segmentations, similar to those in the two HDW-based longitudinal VBM methods. The original BrainWeb GM is then subtracted from each follow-up and the result smoothed. Figure 3.11(d-f) illustrates this process for one subject.

The gold-standard smoothed subtraction images could be entered into an identical one-sample t-test as the actual sets of VBM subtraction images. For reasons discussed in section 3.3.5, we instead use contrast images (the numerator of the t-statistic, here simply equal to the negated mean over the scans), thresholded at different values for visualisation purposes.

Evaluation of segmentation accuracy

While the FEM simulated data provides a form of gold standard for the regional volume changes, it does not provide a direct standard for evaluating each individual segmentation, since the tissue classification used in the atrophy simulation process is itself the result of a potentially inaccurate non-rigid label propagation procedure. To investigate more directly the relative merits of independent segmentation, and the use of high-dimensional warping with pre- or post-averaging, a collection of images with accurate individual segmentations is needed. BrainWeb [26] recently added twenty new anatomical models [27] to their online brain phantom data-base (henceforth referred to as BrainWeb20). Grey-level simulated

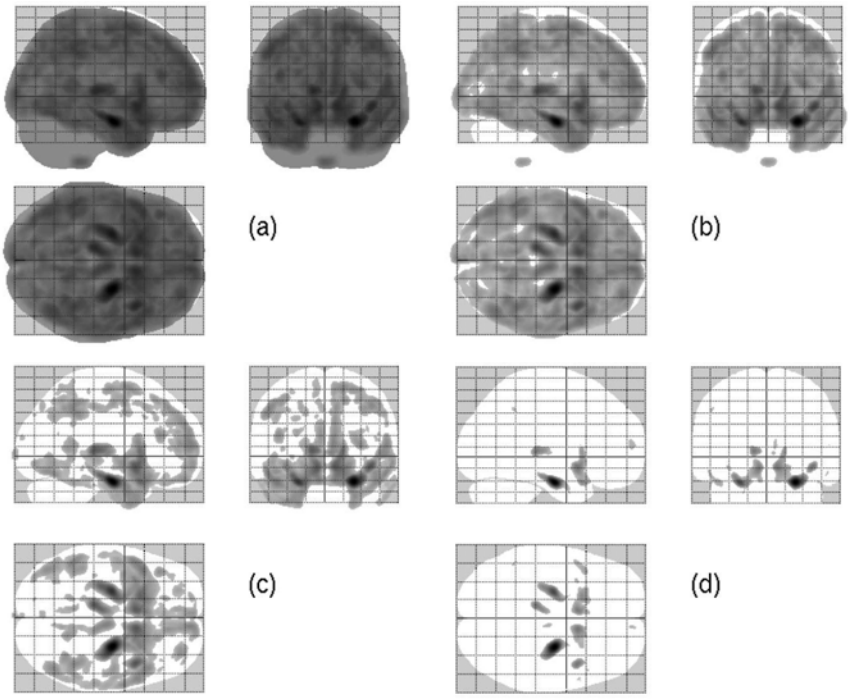


Figure 3.12: Gold-standard average atrophy, Maximum Intensity Projections thresholded at: (a) 0, (b) 0.01, (c) 0.02, (d) 0.03.

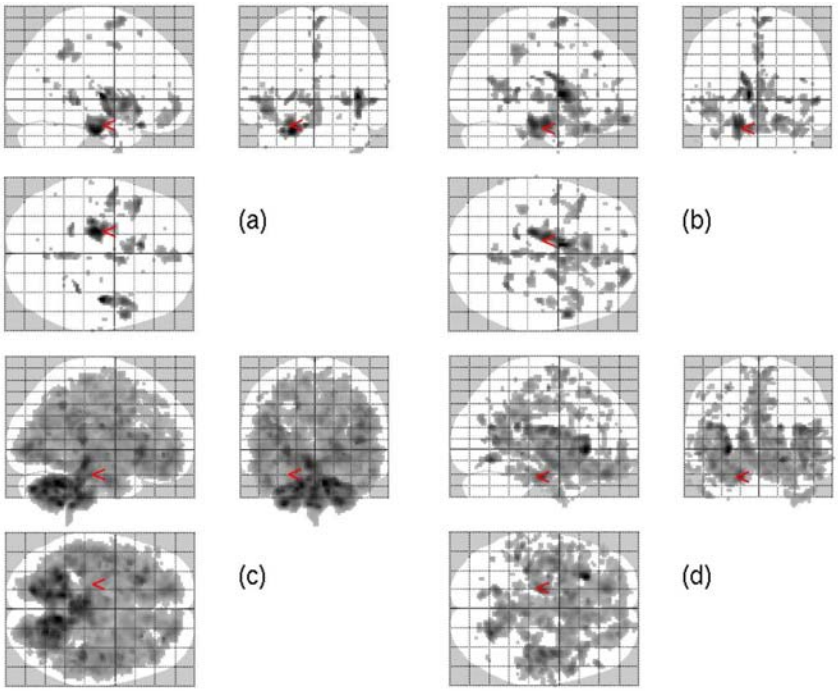


Figure 3.13: Maximum Intensity Projections of significant atrophy ($p_{FWE} < 0.01$) for VBM methods: (a) Standard; (b) Tied-normalisation; (c) Pre-averaged; (d) Post-averaged.

images are available with the underlying tissue model used for their simulation, providing a good gold standard for segmentation.

The HDW-based methods are only meaningful in the context of longitudinal data, which is not provided by BrainWeb. A good solution to this would be to apply the techniques of atrophy simulation described in the previous section, with 'baseline' images from BrainWeb20. The FEM process (and related pre- and post-processing) is time-consuming, and currently somewhat operator-intensive. As a compromise, to provide a simple pilot study, the unrealistic simulation of 'atrophy' with a simple uniform shrinking of the entire image (meaning ventricles, skull, extra-cranial material, etc. all shrink), as used in papers such as [40, 41], has here been applied to just four of the BrainWeb20 images. Rician noise has been added to the baseline and repeat images (after shrinking the latter) using the approach of taking the magnitude after adding complex Gaussian noise [42]. No intensity non-uniformity has thus far been added, though it would be a trivial extension.

3.3.4 Results

FEM simulated atrophy

Gold-standard maximum intensity projections, at varying thresholds, can be seen in figure 3.12. The thresholded images corresponding to figures 3.12(c) and 3.12(d) are also overlaid on coronal sections in figure 3.14.

Statistical results from the four VBM methods are presented in figure 3.13 as maximum intensity projections, and in figure 3.14 as coronal overlays. In both cases the atrophy t-contrast (increase<0) is shown. For the 'reverse contrast' (i.e. gain of GM over time), none of the four methods detected any voxels at the corrected ($p_{FWE} < 0.01$) level.

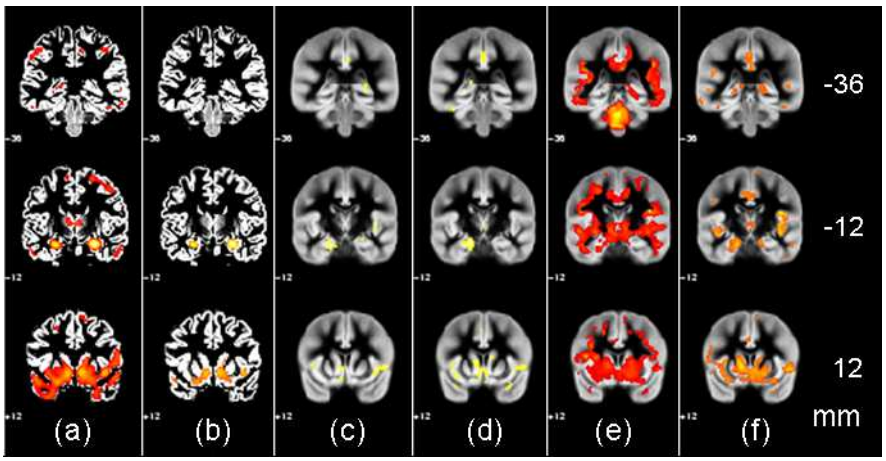


Figure 3.14: Atrophy images overlaid on coronal slices, at indicated positions anterior of the Anterior Commissure: Gold standard contrast-image thresholded at (a) 0.02 and (b) 0.03, on BrainWeb grey-matter. VBM Results, on SPM5 grey-matter tissue probability map: (c) Standard; (d) Tied-normalisation; (e) Pre-averaged; (f) Post-averaged.

Table 3.4 shows correlations between the ground truth contrast image and the contrast or t-value images for the four methods:

Method	std	tied	pre	post
contrast	0.24	0.17	0.63	0.65
t-values	0.16	0.10	0.47	0.36

Table 3.4: Image-wise Spearman rank-correlations (over in-mask voxels) between longitudinal VBM contrast- and t-maps and gold-standard contrast-map.

Segmentation of shrunken BrainWeb20 images

Tissue classification performance of the standard method (independent segmentation) and the two HDW-based longitudinal segmentation methods (pre- and post-averaging) is compared in the following two figures. Two metrics have been used for quantifying the accuracy of the estimated probabilistic segmentations with respect to the fuzzy tissue models underlying the BrainWeb20 simulated images. The fuzzy-overlap measure of Crum et al. [28] has been used directly, and the more conventional Tanimoto overlap (on which the fuzzy-overlap is based) has been used on the binarised segmentations resulting from thresholding at 50% probability. Figure 3.15 shows the results for grey-matter (of primary interest in most VBM studies, including this one), while figure 3.16 shows equivalent results for white-matter.

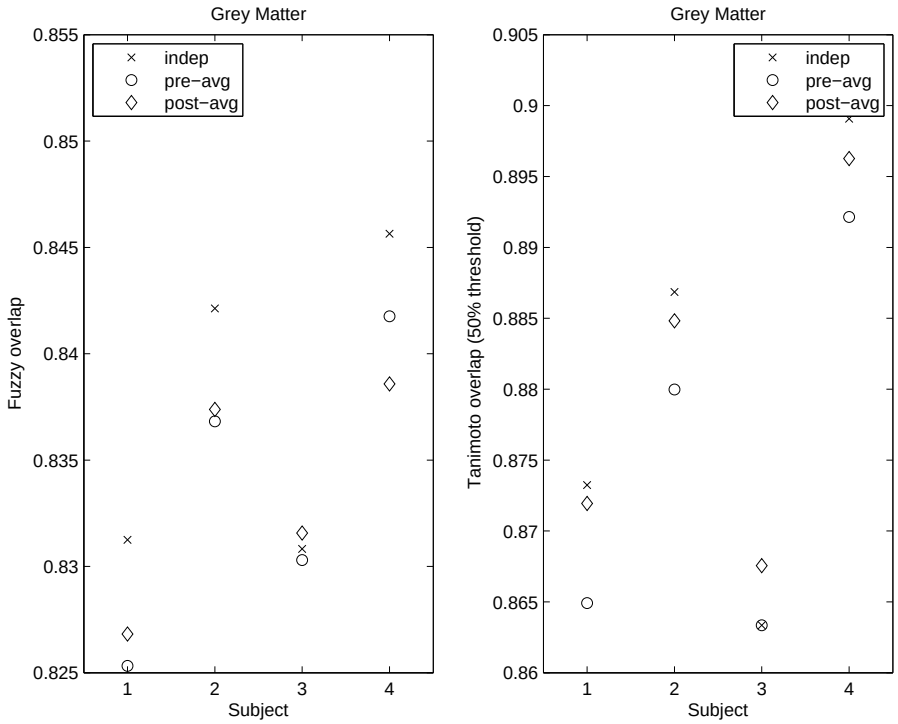


Figure 3.15: Accuracy of grey matter segmentation, quantified with fuzzy overlap and Tanimoto coefficient after binarisation at 50% probability.

3.3.5 Discussion

FEM simulated atrophy

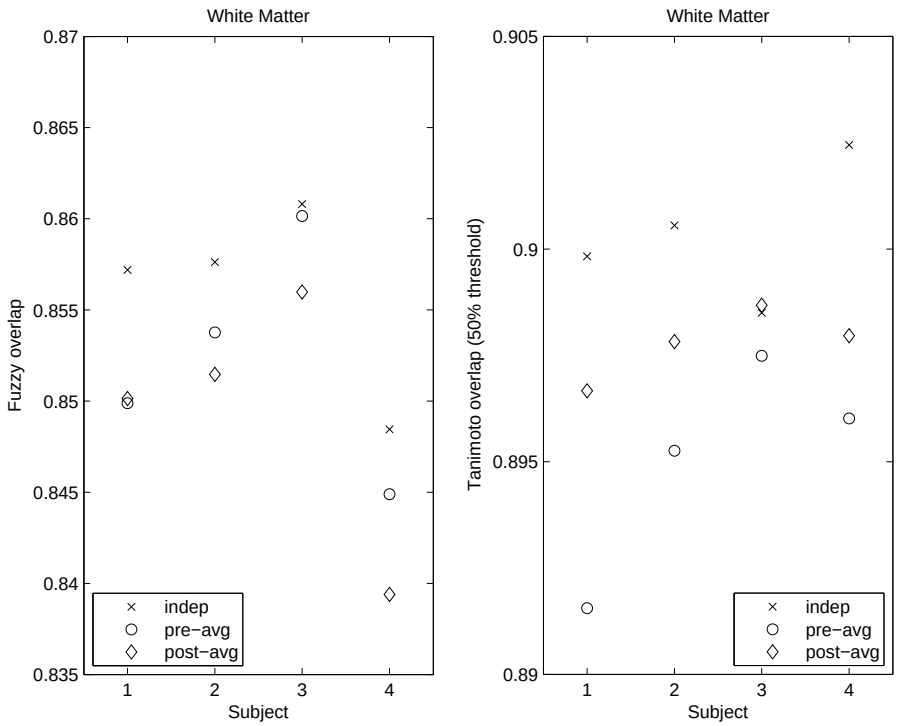


Figure 3.16: Accuracy of white matter segmentation, quantified with fuzzy overlap and Tanimoto coefficient after binarisation at 50% probability.

From Simulation Ground Truth to VBM Gold Standard The atrophy simulation method gives the ground-truth volume change over the nodes of the deformed mesh. However, several steps are required to convert this data into a gold standard for VBM. The volume change maps could probably be improved, for example with mesh-refinement and/or better interpolation from mesh nodes to voxels, but the smoothed GM-increase images which lead to the gold standard appear perfectly adequate.

Following the obvious approach of performing the same statistical analysis of the ground-truth GM-increase images as of the VBM subtraction images, we obtained unrealistic t-maps (not shown). Unreasonably large t-values occur outside the regions in which FEM volume changes were introduced. This may be due to the low spatial variance of the volume change maps outside these areas. Inter-subject spatial variability is far lower for these images than it is for natural anatomical variation in real patient images — even after spatial normalisation. Peak SPM t-values tend to move toward regions of lower spatial variability; this is discussed at length by Reimold et al. [19], who propose a solution based on combining information from the t-map and the contrast image to locate peaks more precisely.

Here, it is desirable to have a gold standard image, instead of a list of peak locations, so we instead threshold the contrast image, in this case simply equal to the (negative) mean GM-increase. The maximum intensity projections and slices (see figures 3.12 and 3.14) now look sensible, but this method leaves open the question of at what level the contrast image should be thresholded, since appeal can no longer be made to statistical

grounds.[12]

The approach taken here of presenting differently-thresholded versions of the gold standard allows a multi-scale evaluation of the pattern of atrophy. Note that the purpose of the gold standard is to indicate the spatial pattern of simulated atrophy; its significance is not of innate interest. To ensure a fair comparison, the statistical thresholds for the different longitudinal methods were held constant.

Longitudinal VBM on simulated atrophy The gold-standard results shown in figures 3.12 and 3.14(a,b) indicate the presence of diffuse global atrophy, with greater focus on the temporal lobes and strongest change in the hippocampi; the cerebellum and brainstem are spared. The key advance of this work is that the VBM methods may be compared directly to this desired pattern, as well as to each other.

We first note that all four methods appear to perform better at detecting the hippocampal and temporal lobe atrophy compared to the more diffuse cortical atrophy. This is probably due to the greater natural anatomical variation in the pattern of cortical folding. Inter-subject registration with accurate sulcal matching is notoriously difficult; some of the most successful approaches require manual intervention [43].

Standard VBM detects the least atrophy of the four methods, though there are no obvious false-positive regions. Longitudinally tied normalisation seems to give only minor improvements, though there is some evidence that the less variable tied registration preserves more of the cortical atrophy.

Both HDW-based methods appear much more sensitive, though some of the atrophy they report is not apparently well-matched to the gold standard (e.g. the insula). In addition, some areas present in the gold standard appear to be missed despite the greater apparent sensitivity (e.g. temporal horns, and the focal nature of the hippocampal atrophy).

The pre-averaging method [34] seems to produce false-positive results in the cerebellum and brainstem. Our new post-averaging method appears to avoid this, at the expense of detecting less true cortical atrophy. Additionally, the post-averaging method has better detected the hippocampal atrophy. Reasons for these differences are not entirely clear, as both methods used the same HDW transformations. The pre-averaging method segments (and normalises) an image with higher signal-to-noise ratio but potentially significant blurring; while post-averaging of original (lower SNR segmentations) also improves the SNR of the results. It is conceivable that false-positives in the cerebellum could arise because the greater blurring in the pre-averaging method causes the balance of GM and WM to change due to the more severe partial volume effect, given the slightly greater bulk of WM versus the thinner and more convoluted layer of GM. The relative merits of these two alternatives need further investigation.

We note that there are statistical objections to the comparison of t- or p-values, as a difference in significance is not equivalent to a significant difference. However, with

[12] We briefly explored the use of permutation-testing on the (non-pivotal) statistic given by the numerator of the usual t-statistic, however, the thresholds turned out to be unhelpfully severe, removing much of the atrophy which we know (from the simulation) should be present.

fundamentally different methods such as the standard and HDW-based VBM evaluated here, there is a risk of registration problems if the ANOVA interactions between atrophy and method are tested. The findings shown here are intended to allow comparison between distributions of detected atrophy of the four methods and the gold standard, with the aims of informing choices between different VBM methods, and guiding further comparative studies.

Segmentation performance

Figures 3.15 & 3.16 show the same general patterns. Surprisingly, given the apparently superior VBM results, the standard independent segmentations appear to be superior to those using high-dimensional image registration. With either metric, and either grey or white matter, standard VBM has the highest segmentation accuracy for at least three of the four subjects. Post-averaging seems marginally superior to pre-averaging based on the Tanimoto overlap of the thresholded segmentations, but pre-averaging appears better judging from the fuzzy-overlap. This inconsistency suggests simply that there is no significant difference between the two methods — an hypothesis which greater amounts of data should be able to confirm or refute.

Performance was also quantified using a third metric: Pearson correlation of the probabilistic segmentations. These results (not shown) exhibited a similar pattern, with the standard method again appearing more successful than either HDW technique, but with little evidence for a significant difference between pre- and post-averaging.

It seems, in this case, that the registration is not sufficiently accurate to be improving the segmentation, even on these images with simple global scaling. In practice, the VBM results do appear to be better for the HDW methods, suggesting that there is still very valuable information in the warps. The obvious conclusion from these findings is that Jacobians from HDW registration to follow-up images should be combined with independently produced baseline segmentations. Interestingly, the method used in Kipps et al. [44] does just this. An earlier version of the software developed during this project did in fact use an SPM5-adapted version of the method of Kipps et al., but after testing (only) on real images, it was dropped in favour of the apparently superior pre-averaging method. Later, the post-averaging method was developed in the hope of improving upon pre-averaging (which it appears to have done). Ironically, it now seems (on the basis of this admittedly limited investigation using unrealistic simulated atrophy) that the simpler approach of Kipps et al. may be the best method for combining HDW Jacobians with segmentations for longitudinal VBM. However, this may well change if a more precise registration method is used for the longitudinal averaging. A groupwise registration method like DARTEL [11] would additionally remove the potential for bias that arises in choosing the baseline as the target for registration of the follow-up images.

Future investigation with greater numbers of subjects (and, ideally, more realistic atrophy simulation, as well as real data) will be essential to provide firm support for conclusions. One appealing idea which we have started to explore, is to quantify preprocessing quality using disease-group classification performance from an automatic machine-learning

classifier, such as a support vector machine (SVM), as used by Klöppel et al. [45]. The main challenge here is that large data sets are required for precise and reliable estimates of classification accuracy, but only relatively small data sets are available with completely certain classification ground truth. Atrophy simulation may have a part to play here too.

3.4 Guidelines for reporting VBM studies

3.4.1 Abstract

Voxel-Based Morphometry [4] is a commonly used tool for studying neuroanatomical correlates of subject characteristics and patterns of brain change in development or disease. In performing a VBM study, many methodological options are available; if the study is to be easily interpretable and repeatable, then processing steps and decisions must be clearly described. Similarly, unusual methods and parameter choices should be justified in order to aid readers in judging the importance of such options or in comparing the work with other studies. In this section, we suggest core principles that should be followed and information that should be included when reporting a VBM study, in order to make it transparent, replicable and useful.

3.4.2 Introduction

Voxel-Based Morphometry [4, 46] is becoming increasingly widely used as a tool to examine patterns of brain change in healthy ageing [36] or neurodegenerative disease [47] and neuroanatomical correlates of behavioural or cognitive deficits [48]. VBM essentially involves voxel-wise statistical analysis of preprocessed structural MR images. Although much of the processing and analysis is automated in software packages such as SPM, many methodological decisions remain, ranging from what template to use for normalisation, to what level and type of correction to use and how best to display results. Different approaches, such as VBM using RAVENS maps [10], introduce yet more options. It can be difficult to replicate or draw conclusions from VBM studies if the processing steps are not clearly described. Similarly, if unusual methods or parameters are employed without sufficient justification it can be challenging for readers to judge the potential impact on results or to compare the work with other studies. Here, we present a set of recommendations, in the form of ten 'rules', which we hope will be helpful to authors when writing up VBM studies. The rules are intended to outline core principles that should be followed and information that should be included when reporting a VBM study, in order to make it transparent, replicable and useful. Since the field is rapidly developing, such rules must not be overly restrictive; some of the points below aim to explain more general principles, in the hope of aiding the reader to follow good practice in areas where rigid protocols cannot be given. We feel that guidelines are crucial for clear scientific communication and the development of the field, as VBM data sets accumulate and alternative procedures and techniques proliferate. Additional motivation for this work came from the success of the CONSORT statement [49], a major undertaking that helped to standardise and improve the reporting of randomised controlled trials, and from a related effort in the field of functional brain imaging [50].[13]

[13] See also: `http://www.fmrimethods.org`

3.4.3 Rules

1. Set out the rationale for your study and describe the data fully

What are the key experimental questions, and why was VBM preferred over other techniques in order to address these questions? Prior hypotheses should be stated; either experimental ones or a priori anatomical or spatial regions in which you predict effects might be found [51]. This is particularly important if search volumes are restricted when correcting for multiple statistical tests during data analysis (see Rule 5). The study design should be described in enough detail for readers to be confident that subjects have been included appropriately and that important sources of error have been identified, and, where possible, controlled for. Subject inclusion and exclusion criteria should be clearly set out, as well as baseline demographic information (such as age, gender and handedness) and any other variables which are relevant to the interpretation of the findings [52]. Examples of such variables could include IQ in a study of cognitive function, or measures of disease severity or duration in a clinical study. Image acquisition can influence morphometry results [53]; it is therefore essential to report any variations in acquisition, such as different scanners, scanner upgrades, or pulse sequence changes. The relative timing of data acquisition should be specified, for example, were MRI and any other clinical or behavioural data for a subject collected on the same day; if not what was the interval? It is also important to specify whether MRI data for different groups were collected in an interleaved fashion, or in blocks (thus raising the possibility that changes in scanner calibration over time could be an additional source of inter-group variation, [54]). Scanner locations and make should be mentioned for multi-centre studies, and assessment interval (for MRI and any other data collection) should be made clear for longitudinal studies.. If analysing multiple groups (e.g. patients and controls), discuss whether potential confounds, such as age, gender or acquisition differences, are balanced between groups. If subjects or scans were excluded from the analysis this should be stated and justified (see Rule 9).

2. Explain how the brain segmentations are produced

The inputs to VBM's statistical analysis are derived from structural MR images using tissue-segmentation, spatial normalisation, and smoothing. Additional preprocessing is often performed before the main segmentation step, generally using automatic algorithms such as MR bias correction or skull-stripping, or manual techniques such as semi-automatic brain-segmentation or interactive reorientation. Multiple processes may be combined within unified algorithms, such as that of Ashburner and Friston [30]. The preprocessing steps must be reported in sufficient detail for the methods to be clear and reproducible; as a minimum, this should include the software packages used (with version numbers) and any parameters altered from the default values. For interactive steps, authors should clarify the protocol, for example whether operators were blind to subject identity. The segmentation method itself should be reported so as to be reproducible; either through clear identification of the software package and description of any defaults modified, or

via careful description of the algorithm. Some popular segmentation algorithms use registered spatial priors, in which case the source of the priors and the means of alignment should be clear. In particular, with SPM2, different methods of iterative segmentation and normalisation have been used, often including iterative regeneration of priors [36, 55]; these should be reported in detail — terms such as 'optimised VBM using SPM2' are not sufficiently precise. Following segmentation, other image-processing methods can be used to condition the data further. Such techniques include morphological filtering (used in the 'clean-up' option of SPM2 and 5), the application of Markov Random Field models,[14] or interactive editing of segmentations. These approaches tend to be less standardised, so should be reported carefully. The final image-processing step is usually to smooth the segmentations, typically through convolution with a Gaussian kernel, in which case the Full-Width at Half-Maximum (FWHM) should be reported. Since smoothing sensitises the analysis to a particular spatial scale of effect (due to the matched filter theorem [13]) some justification of the choice of FWHM may be helpful. Less widely used smoothing techniques, such as anisotropic smoothing [56], should be explained in detail.

3. Describe the method of inter-subject spatial normalisation

In order to compare different subjects, it is essential to use some kind of registration algorithm to bring the images into at least approximate correspondence. Both the technique used and the reference space to which brains are aligned can impact on the results [57], so clear reporting is crucial. As with the other preprocessing steps (see Rule 2), if a popular software package is used, deviations from the default options should be highlighted. The less standard the approach, the more detailed the description should be — as a minimum it should include the four basic elements of image registration: the spatial transformation model; the objective function, including any regularisation terms or Bayesian priors; the optimisation algorithm; and the interpolation method. Spatially-normalised segmentations may be subsequently 'modulated' with the Jacobian determinants from the transformation, in order to adjust for the resulting volume changes. This can heavily influence the results and their interpretation [46, 58], so authors should state whether or not modulation has been performed and justify this choice (particularly if non-rigid registration is used without modulation). It is important to clearly report the reference space to which brains are being aligned, as there are a number of different options available that are defined in quite different ways; ranging from low degree of freedom landmark based reorientation and scaling [59], to automated registration with greater degrees of freedom, either to a template [60, 61] or to tissue probability maps [30]. Template images or segmentations may be standard, such as the popular MNI or ICBM ones used in SPM and FSL, or may be derived from the subjects themselves [11, 36, 62, 63]. If a subset of the data are used to generate custom templates or tissue probability maps, then which subjects (e.g. healthy, diseased, or a balanced mix), and why, should be clear. Poldrack et al. [50] further discuss the choice of reference space, with particular focus on the concept of Talairach space and its relation to standard atlases.

[14] See e.g. Christian Gaser's software at http://dbm.neuro.uni-jena.de/vbm/markov-random-fields/

4. Make your statistical design transparent

There are two issues here, model specification, and contrast testing. When constructing a model it is important to be clear about which variables are included, and why. In the case of factorial designs, it should be obvious to the reader exactly what the factors were, the levels of each factor, and which interactions between factors were modelled. With estimation methods more advanced than Ordinary Least Squares, it may be necessary to report extra information; for example, SPM5 includes non-sphericity options that allow levels of a factor to be dependent or to have different variances. Subject characteristics (Rule 2) should be assessed critically to ensure confounding variables have been included as covariates where appropriate. It is helpful to the reader to indicate why each variable has been modelled, and whether it is a variable of interest (e.g. a psychological score) or a potentially confounding factor (e.g. age). It may also be desirable to adjust for each subject's total or global brain tissue-volume, integrated over the whole analysis search region; either by entering the global values as a covariate, or using them to scale the original voxel values (See Kiebel and Holmes, chapter 8 of [64], for a tutorial in the context of PET imaging). Adjusting for total intra-cranial volume [54] should also be considered, as it can affect the results [36]. Adjustment for global variables remains a topic of debate in VBM [46]. For all covariates, options relating to centring or orthogonalisation should be reported, especially if factor-covariate interactions are modelled. When interrogating the model, the contrasts tested should be described precisely, in terms of the variables involved and their weights. The choice of statistic (t-test or F-test) should be justified and (for single-tailed t-tests) the direction specified. Inclusion of a diagram (e.g. the design matrix) or equation summarising the model and contrasts may be helpful.

5. Be clear about the significance of your findings

As with other mass-univariate image analysis techniques, VBM performs a very large number of statistical tests. The method used to correct for multiple testing should be both clearly stated, and carefully considered — ideally, a priori. VBM is often performed on limited numbers of subjects (for example, to investigate rare disorders), where there is a temptation to report uncorrected results due to low statistical power; this should be made obvious, if done, and is probably best avoided — alternatives include correction at a less stringent alpha level (it is then clear that there has been an effort to control the false positive rate, and to what extent), or clear presentation of unthresholded t- or effect-maps.[15] Studies have also been published comparing single subjects to larger control groups; the standard parametric statistical framework may be poorly suited to such unbalanced designs unless large smoothing kernels are employed [65]. Control of the voxel-level Family-wise Error rate (FWE) using methods based on random field theory requires estimation of the smoothness of the data, and depends strongly on the size of the search region. Therefore, interpretation would be aided by reporting the estimated FWHM smoothness (not the same as the smoothness applied during preprocessing) and

[15] See Brett's comments at `http://imaging.mrc-cbu.cam.ac.uk/imaging/UnthresholdedEffectMaps`.

the resel count. In addition, the method used to define the search region (e.g. an explicit mask, or an absolute or relative threshold) should be specified. Cluster-level control of FWE usually assumes stationary smoothness, which is unlikely to be appropriate for VBM, unless special techniques are employed;[16] if used it should be justified, and the cluster-defining threshold must be reported. Permutation-based statistics [16] provide an alternative method to control FWE (based on voxel value, cluster-size or cluster-mass). These make fewer assumptions, but require careful explanation of the statistical design (see chapter 2). If sub-volumes of the main search region are analysed (known as Small Volume Correction in SPM) authors should explain how and why these regions of interest were selected. Such regions should ideally be anatomically-defined and chosen a priori with justification. (see also Rule 8). False Discovery Rate correction (FDR) [18] can follow either parametric or permutation-based statistics, over the whole search region or sub-volumes, and recently Chumbley and Friston [66] have suggested applying FDR to uncorrected cluster-based p-values. These choices mean reporting should be more detailed than a simple statement that FDR was used.

6. Present results unambiguously

The type and level of correction should be stated in all figure and table legends, and if the statistical parametric map (SPM) is displayed as orthogonal slices or sections then coordinates should be given. It may be helpful for tables to include statistic values and cluster sizes, as well as coordinates of local maxima. SPMs should be displayed on a template that represents some form of average anatomy, for example, the MNI T1-template often used for normalisation, or ideally, a study-specific mean image. Displaying overlays on a single high-resolution image is misleading: an individual subject is likely to be poorly representative of the group, and implies a higher level of anatomical precision than is possible with smoothed data [67].[17] A similar caveat applies to the use of anatomical labels. Methods for converting MNI coordinates to Talairach space should be referenced,[18] and may be best avoided [67]. Comparison of results can be aided by using the same t- or F-statistic colour-scales across figures. If an SPM is displayed at a threshold lower than that used to locate significant voxels (for example in order to show small effects or give an impression of the overall distribution of change) this should be made explicit. If single-tailed t-tests are focussed on (for example in a study of atrophy where tissue gain would be clinically implausible), it may nevertheless be helpful to report the reverse contrast; it can indicate misregistration as a potential confound — or even a possible cause — for the main findings.

7. Clarify and justify any non-standard statistical analyses

As a general principle, the less standard the analysis, the more thoroughly it should be explained (see also Rule 3). Here, we discuss three of the more common examples.

[16] Such as Satoru Hayasaka's toolbox for SPM — http://fmri.wfubmc.edu/cms/NS-General

[17] Consider also the limitations of spatial normalisation discussed in Rule 9.

[18] For further discussion of the concept of Talairach space, see Poldrack et al. [50].

Contrast masking may be used to disambiguate multiple possible causes of an effect or to define smaller search regions, in which case authors should clarify not only which contrasts were analysed and which were used for masking (and at what threshold), but also the motivation for doing so and their interpretation. If a conjunction of analyses is tested using the minimum of several statistic images, it is crucial to clarify the null hypothesis — global, conjunction, or intermediate [68]. If data are extracted (e.g. eigenvariates from volumes of interest, peak voxels or cluster summaries) for analysis with other statistical software, this should be explained and justified (see also the Rule below).

8. Guard against common pitfalls

Here we discuss a few potential problems with VBM analyses that might be easily overlooked. Firstly, note that while voxel-wise multiple testing is usually corrected for (see Rule 5), most software packages do nothing to correct for the user's investigation of multiple contrasts — the more conventional multiple-comparison problem [69]. A simple example of this could occur if two opposite single-tailed t-contrasts were analysed; if any findings in either could be reported as significant, then the alpha-level or p-values should be adjusted to reflect this — even if the other analysis is not reported. With more complex models it can be difficult to decide on a suitable correction procedure [70], but if many contrasts have been tested and not reported, this should be noted. A more insidious multiple-comparisons problem can occur if part or all of the VBM analysis is repeated for any reason. The motivation for this is crucial: for example, different amounts of smoothing (see Rules 2 and 9) may be used to match the filter size to multiple spatial scales of expected effects, whereas it would be misleading to try several FWHM values before reporting only the most appealing results. It is also possible to invalidate correction for voxel-wise multiple tests by extracting sub-regions of the images for further analysis; it is essential that the procedure used to select data is independent of the subsequent analysis [71], and clearly described. Similar caveats apply to the selection of alternative parameters at other preprocessing stages, or the analysis of multiple sub-groups of subjects (perhaps for disease sub-types), unless this is done using independent data sets. It is sometimes necessary to exclude certain subjects or scans (for example due to artefacts or preprocessing failures) such decisions should ideally be blind to the subjects' identity, and care should be taken to avoid bias or, if this is not possible (e.g. if more severely affected subjects are more likely to be excluded due to poor segmentation), sources of bias should be acknowledged.

9. Recognise the limitations of the technique

Like all image analysis methods, VBM has inherent limitations [72]. The basic premise of inter-subject spatial normalisation is problematic: different subjects can have different gyral variants with no 'true' correspondence between them; and information from structural MRI (even manual sulcal labelling) does not necessarily predict underlying cytoarchitectonic borders [73]. Normalisation accuracy is also likely to vary between brain regions, for example highly convoluted cortex will register less well than simpler structures. This

suggests that conclusions regarding fine-scale anatomical localisation should be cautious; there is no single 'correct' normalisation method. Smoothing can alleviate some of the problems of inter-subject correspondence (in addition to making the data more normally distributed) but brings problems of its own. Variations in smoothing can produce very different results [74], and while investigators may have a rough idea of a reasonable kernel size for their study (based on a priori beliefs about the likely scale of interest), a degree of arbitrariness remains. All classical statistical tests share the limitation that failure to reject the null-hypothesis does not imply that it is true (this is particularly pertinent if tests only just fail to reach arbitrary significance levels, such as 0.05). More specifically, with SPM, the absence of a statistically significant effect in a particular region does not prove that the region is unaffected. This is especially true for VBM, where regional variation in normalisation accuracy [75] or smoothness [4] is likely to cause spatially variable sensitivity.

10. Interpret your results cautiously and in context

When implemented rigorously and interpreted carefully VBM can be a powerful technique. Authors should be forthright in discussing potential sources of bias or imprecision, whether they arise from the study's design or analysis, or from the nature of VBM itself. Particular care should be taken when interpreting results which appear fragile with respect to more arbitrary aspects of the method such as preprocessing options and nuisance variables. A conservative approach based on robust findings, related to a priori hypotheses, is preferable to reporting weak effects that may be idiosyncratic to the particular parameters chosen. This approach reflects an awareness of the potential sources of error and bias that can be introduced at the different stages of a VBM study — effects that are likely to be amplified in clinical populations with inherently atypical anatomy. Despite the caveats, our basic message is brief: your VBM study should be conducted and reported in a way that is principled, transparent and replicable. Such studies have potential to become valuable contributions to the literature.

Bibliography

[1] O. Camara, M. Schweiger, R. Scahill, W. Crum, B. Sneller, J. Schnabel, G. Ridgway, D. Cash, D. Hill, and N. Fox, "Phenomenological model of diffuse global and regional atrophy using finite-element methods," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1417–1430, Nov. 2006. ˆ145, 161, 163

[2] G. R. Ridgway, S. M. D. Henley, J. D. Rohrer, R. I. Scahill, J. D. Warren, and N. C. Fox, "Ten simple rules for reporting voxel-based morphometry studies." *Neuroimage*, vol. 40, no. 4, pp. 1429–1435, May 2008. ˆ145, 159

[3] I. C. Wright, P. K. McGuire, J. B. Poline, J. M. Travere, R. M. Murray, C. D. Frith, R. S. Frackowiak, and K. J. Friston, "A voxel-based method for the statistical analysis

of gray and white matter density applied to schizophrenia." *Neuroimage*, vol. 2, no. 4, pp. 244–252, Dec. 1995. ˆ145

[4] J. Ashburner and K. J. Friston, "Voxel-based morphometry–the methods." *Neuroimage*, vol. 11, no. 6 Pt 1, pp. 805–821, Jun. 2000. ˆ145, 146, 172, 178

[5] M. Miller, A. Banerjee, G. Christensen, S. Joshi, N. Khaneja, U. Grenander, and L. Matejic, "Statistical methods in computational anatomy." *Stat Methods Med Res*, vol. 6, no. 3, pp. 267–299, Sep. 1997. ˆ145

[6] U. Grenander and M. I. Miller, "Computational anatomy: an emerging discipline," *Quarterly of Applied Mathematics*, vol. 56, no. 4, pp. 617–694, 1998. ˆ145

[7] J. Ashburner, "Computational neuroanatomy," Ph.D. dissertation, University College London, 2000. [Online]. Available: http://www.fil.ion.ucl.ac.uk/spm/doc/theses/john/ ˆ145

[8] M. I. Miller, G. E. Christensen, Y. Amit, and U. Grenander, "Mathematical textbook of deformable neuroanatomies." *Proc Natl Acad Sci U S A*, vol. 90, no. 24, pp. 11 944–11 948, Dec. 1993. ˆ145

[9] C. D. Good, J. Ashburner, and R. S. Frackowiak, "Computational neuroanatomy: new perspectives for neuroradiology." *Rev Neurol (Paris)*, vol. 157, no. 8-9 Pt 1, pp. 797–806, Sep. 2001. ˆ145

[10] C. Davatzikos, A. Genc, D. Xu, and S. M. Resnick, "Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy." *Neuroimage*, vol. 14, no. 6, pp. 1361–1369, Dec. 2001. ˆ146, 158, 172

[11] J. Ashburner, "A fast diffeomorphic image registration algorithm." *Neuroimage*, vol. 38, no. 1, pp. 95–113, Oct. 2007. ˆ146, 158, 170, 174

[12] J. Ashburner and K. J. Friston, "Computing average shaped tissue probability templates." *Neuroimage*, vol. 45, no. 2, pp. 333–341, Apr. 2009. ˆ146

[13] ——, "Why voxel-based morphometry should be used." *Neuroimage*, vol. 14, no. 6, pp. 1238–1243, Dec. 2001. ˆ146, 174

[14] C. H. Salmond, J. Ashburner, F. Vargha-Khadem, A. Connelly, D. G. Gadian, and K. J. Friston, "Distributional assumptions in voxel-based morphometry." *Neuroimage*, vol. 17, no. 2, pp. 1027–1030, Oct. 2002. ˆ146

[15] K. Worsley, S. Marrett, P. Neelin, and A. Evans, "Searching scale space for activation in PET images," *Human Brain Mapping*, vol. 4, no. 1, pp. 74–90, 1996. ˆ146, 158

[16] T. E. Nichols and A. P. Holmes, "Nonparametric permutation tests for functional neuroimaging: A primer with examples," *Human Brain Mapping*, vol. 15, no. 1, pp. 1–25, 2002. ˆ147, 176

[17] M. Belmonte and D. Yurgelun-Todd, "Permutation testing made practical for functional magnetic resonance image analysis," *IEEE Trans. Med. Imag.*, vol. 20, no. 3, pp. 243–248, Mar. 2001. ˆ147

[18] C. R. Genovese, N. A. Lazar, and T. Nichols, "Thresholding of statistical maps in functional neuroimaging using the false discovery rate." *Neuroimage*, vol. 15, no. 4, pp. 870–878, Apr. 2002. ˆ147, 176

[19] M. Reimold, M. Slifstein, A. Heinz, W. Mueller-Schauenburg, and R. Bares, "Effect of spatial smoothing on t-maps: arguments for going back from t-maps to masked contrast images." *J Cereb Blood Flow Metab*, vol. 26, no. 6, pp. 751–759, Jun. 2006. ˆ147, 168

[20] Z. Lao, D. Shen, Z. Xue, B. Karacali, S. M. Resnick, and C. Davatzikos, "Morphological classification of brains via high-dimensional shape transformations and machine learning methods." *Neuroimage*, vol. 21, no. 1, pp. 46–57, Jan. 2004. ˆ147

[21] P. Vemuri, J. L. Gunter, M. L. Senjem, J. L. Whitwell, K. Kantarci, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack, "Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies." *Neuroimage*, vol. 39, no. 3, pp. 1186–1197, Feb. 2008. ˆ147, 148

[22] K. Friston, C. Chu, J. Mourão-Miranda, O. Hulme, G. Rees, W. Penny, and J. Ashburner, "Bayesian decoding of brain images." *Neuroimage*, vol. 39, no. 1, pp. 181–205, Jan. 2008. ˆ147

[23] G. B. Karas, E. J. Burton, S. A. R. B. Rombouts, R. A. van Schijndel, J. T. O'Brien, P. Scheltens, I. G. McKeith, D. Williams, C. Ballard, and F. Barkhof, "A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry." *Neuroimage*, vol. 18, no. 4, pp. 895–907, Apr. 2003. ˆ147

[24] C. J. Mummery, K. Patterson, C. J. Price, J. Ashburner, R. S. Frackowiak, and J. R. Hodges, "A voxel-based morphometry study of semantic dementia: relationship between temporal lobe atrophy and semantic memory." *Ann Neurol*, vol. 47, no. 1, pp. 36–45, Jan. 2000. ˆ147

[25] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. J. Behrens, H. Johansen-Berg, P. R. Bannister, M. D. Luca, I. Drobnjak, D. E. Flitney, R. K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. D. Stefano, J. M. Brady, and P. M. Matthews, "Advances in functional and structural MR image analysis and implementation as FSL." *Neuroimage*, vol. 23 Suppl 1, pp. S208–S219, 2004. ˆ148

[26] C. Cocosco, V. Kollokian, R. Kwan, and A. Evans, "Brainweb: Online interface to a 3D MRI simulated brain database," *NeuroImage*, vol. 5, no. 4, p. S425, 1997. [Online]. Available: http://www.bic.mni.mcgill.ca/users/crisco/HBM97_abs/HBM97_abs.html ˆ148, 164

[27] B. Aubert-Broche, M. Griffin, G. Pike, A. Evans, and D. Collins, "Twenty new digital brain phantoms for creation of validation image data bases," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1410–1416, Nov. 2006. ˆ148, 164

[28] W. Crum, O. Camara, and D. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1451–1461, Nov. 2006. ˆ148, 167

[29] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""Mini-mental state". A practical method for grading the cognitive state of patients for the clinician." *J Psychiatr Res*, vol. 12, no. 3, pp. 189–198, Nov. 1975. ˆ148

[30] J. Ashburner and K. J. Friston, "Unified segmentation." *Neuroimage*, vol. 26, no. 3, pp. 839–851, Jul. 2005. ˆ149, 162, 173, 174

[31] K. Worsley, S. Marrett, P. Neelin, A. Vandal, K. Friston, A. Evans *et al.*, "A unified statistical approach for determining significant signals in images of cerebral activation," *Human Brain Mapping*, vol. 4, no. 1, pp. 58–73, 1996. ˆ156

[32] D. Glaser and K. Friston, *Variance Components*, 2nd ed. Academic Press, 2004, ch. 9. [Online]. Available: http://www.fil.ion.ucl.ac.uk/spm/doc/books/hbf2/pdfs/Ch9.pdf ˆ158

[33] B. Draganski, C. Gaser, V. Busch, G. Schuierer, U. Bogdahn, and A. May, "Neuroplasticity: changes in grey matter induced by training." *Nature*, vol. 427, no. 6972, pp. 311–312, Jan. 2004. ˆ161, 162

[34] G. Chételat, B. Landeau, F. Eustache, F. Mézenge, F. Viader, V. de la Sayette, B. Desgranges, and J.-C. Baron, "Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study." *Neuroimage*, vol. 27, no. 4, pp. 934–946, Oct. 2005. ˆ161, 163, 169

[35] J. M. Schott, S. L. Price, C. Frost, J. L. Whitwell, M. N. Rossor, and N. C. Fox, "Measuring atrophy in Alzheimer disease: a serial MRI study over 6 and 12 months." *Neurology*, vol. 65, no. 1, pp. 119–124, Jul. 2005. ˆ161, 163

[36] C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. Henson, K. J. Friston, and R. S. Frackowiak, "A voxel-based morphometric study of ageing in 465 normal adult human brains." *Neuroimage*, vol. 14, no. 1 Pt 1, pp. 21–36, Jul. 2001. ˆ162, 172, 174, 175

[37] N. C. Fox and J. M. Schott, "Imaging cerebral atrophy: normal ageing to Alzheimer's disease." *Lancet*, vol. 363, no. 9406, pp. 392–394, Jan. 2004. ˆ162

[38] D. Collins, A. Zijdenbos, V. Kollokian, J. Sled, N. Kabani, C. Holmes, and A. Evans, "Design and construction of a realistic digital brain phantom," *IEEE Trans. Med. Imag.*, vol. 17, no. 3, pp. 463–468, Jun. 1998. ˆ163

[39] W. R. Crum, C. Tanner, and D. J. Hawkes, "Anisotropic multi-scale fluid registration: evaluation in magnetic resonance breast imaging." *Phys Med Biol*, vol. 50, no. 21, pp. 5153–5174, Nov. 2005. ˆ163

[40] E. B. Lewis and N. C. Fox, "Correction of differential intensity inhomogeneity in longitudinal MR images." *Neuroimage*, vol. 23, no. 1, pp. 75–83, Sep. 2004. ˆ166

[41] R. G. Boyes, D. Rueckert, P. Aljabar, J. Whitwell, J. M. Schott, D. L. G. Hill, and N. C. Fox, "Cerebral atrophy measurements using Jacobian integration: comparison with the boundary shift integral." *Neuroimage*, vol. 32, no. 1, pp. 159–169, Aug. 2006. ˆ166

[42] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data." *IEEE Trans. Med. Imag.*, vol. 17, no. 1, pp. 87–97, Feb. 1998. ˆ166

[43] P. M. Thompson, K. M. Hayashi, E. R. Sowell, N. Gogtay, J. N. Giedd, J. L. Rapoport, G. I. de Zubicaray, A. L. Janke, S. E. Rose, J. Semple, D. M. Doddrell, Y. Wang, T. G. M. van Erp, T. D. Cannon, and A. W. Toga, "Mapping cortical change in Alzheimer's disease, brain development, and schizophrenia." *Neuroimage*, vol. 23 Suppl 1, pp. S2–18, 2004. ˆ169

[44] C. M. Kipps, A. J. Duggins, N. Mahant, L. Gomes, J. Ashburner, and E. A. McCusker, "Progression of structural neuropathology in preclinical Huntington's disease: a tensor based morphometry study." *J Neurol Neurosurg Psychiatry*, vol. 76, no. 5, pp. 650–655, May 2005. ˆ170

[45] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack, J. Ashburner, and R. S. J. Frackowiak, "Automatic classification of MR scans in Alzheimer's disease." *Brain*, vol. 131, no. 3, pp. 681–689, Mar. 2008. ˆ171

[46] A. Mechelli, C. J. Price, K. J. Friston, and J. Ashburner, "Voxel-based morphometry of the human brain: Methods and applications," *Current Medical Imaging Reviews*, vol. 1, no. 1, pp. 1–9, 2005. ˆ172, 174, 175

[47] J. C. Baron, G. Chételat, B. Desgranges, G. Perchey, B. Landeau, V. de la Sayette, and F. Eustache, "In vivo mapping of gray matter loss with voxel-based morphometry in mild Alzheimer's disease." *Neuroimage*, vol. 14, no. 2, pp. 298–309, Aug. 2001. ˆ172

[48] F. Abell, M. Krams, J. Ashburner, R. Passingham, K. Friston, R. Frackowiak, F. Happé, C. Frith, and U. Frith, "The neuroanatomy of autism: a voxel-based whole brain analysis of structural scans." *Neuroreport*, vol. 10, no. 8, pp. 1647–1651, Jun. 1999. ˆ172

[49] D. G. Altman, "Better reporting of randomised controlled trials: the CONSORT statement." *BMJ*, vol. 313, no. 7057, pp. 570–571, Sep. 1996. [Online]. Available: http://www.bmj.com/cgi/content/extract/313/7057/570 ˆ172

[50] R. A. Poldrack, P. C. Fletcher, R. N. Henson, K. J. Worsley, M. Brett, and T. E. Nichols, "Guidelines for reporting an fMRI study." *Neuroimage*, Dec. 2007. ˆ172, 174, 176

[51] E. A. Maguire, D. G. Gadian, I. S. Johnsrude, C. D. Good, J. Ashburner, R. S. Frackowiak, and C. D. Frith, "Navigation-related structural change in the hippocampi of taxi drivers." *Proc Natl Acad Sci U S A*, vol. 97, no. 8, pp. 4398–4403, Apr. 2000. ˆ173

[52] R. I. Scahill, C. Frost, R. Jenkins, J. L. Whitwell, M. N. Rossor, and N. C. Fox, "A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging." *Arch Neurol*, vol. 60, no. 7, pp. 989–994, Jul. 2003. ˆ173

[53] A. Littmann, J. Guehring, C. Buechel, and H.-S. Stiehl, "Acquisition-related morphological variability in structural MRI." *Acad Radiol*, vol. 13, no. 9, pp. 1055–1061, Sep. 2006. ˆ173

[54] J. L. Whitwell, W. R. Crum, H. C. Watt, and N. C. Fox, "Normalization of cerebral volumes by use of intracranial volume: implications for longitudinal quantitative MR imaging." *AJNR Am J Neuroradiol*, vol. 22, no. 8, pp. 1483–1489, Sep. 2001. [Online]. Available: http://www.ajnr.org/cgi/content/abstract/22/8/1483 ˆ173, 175

[55] G. Douaud, V. Gaura, M.-J. Ribeiro, F. Lethimonnier, R. Maroy, C. Verny, P. Krystkowiak, P. Damier, A.-C. Bachoud-Levi, P. Hantraye, and P. Remy, "Distribution of grey matter atrophy in Huntington's disease patients: a combined ROI-based and voxel-based morphometric study." *Neuroimage*, vol. 32, no. 4, pp. 1562–1575, Oct. 2006. ˆ174

[56] G. Gerig, O. Kubler, R. Kikinis, and F. Jolesz, "Nonlinear anisotropic filtering of MRI data," *IEEE Trans. Med. Imag.*, vol. 11, no. 2, pp. 221–232, Jun. 1992. ˆ174

[57] M. L. Senjem, J. L. Gunter, M. M. Shiung, R. C. Petersen, and C. R. Jack, "Comparison of different methodological implementations of voxel-based morphometry in neurodegenerative disease." *Neuroimage*, vol. 26, no. 2, pp. 600–608, Jun. 2005. ˆ174

[58] S. S. Keller, M. Wilke, U. C. Wieshmann, V. A. Sluming, and N. Roberts, "Comparison of standard and optimized voxel-based morphometry for analysis of brain changes associated with temporal lobe epilepsy." *Neuroimage*, vol. 23, no. 3, pp. 860–868, Nov. 2004. ˆ174

[59] J. Talairach and P. Tournoux, *Co-Planar Stereotaxic Atlas of the Human Brain: 3-dimensional Proportional System: an Approach to Cerebral Imaging.* Thieme, 1988. ˆ174

[60] J. Ashburner and K. J. Friston, "Nonlinear spatial normalization using basis functions." *Hum Brain Mapp*, vol. 7, no. 4, pp. 254–266, 1999. ˆ174

[61] D. Shen and C. Davatzikos, "HAMMER: hierarchical attribute matching mechanism for elastic registration." *IEEE Trans. Med. Imag.*, vol. 21, no. 11, pp. 1421–1439, Nov. 2002. ˆ174

[62] P. Kochunov, J. L. Lancaster, P. Thompson, R. Woods, J. Mazziotta, J. Hardies, and P. Fox, "Regional spatial normalization: toward an optimal target." *J Comput Assist Tomogr*, vol. 25, no. 5, pp. 805–816, 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11584245 ˆ174

[63] S. Joshi, B. Davis, M. Jomier, and G. Gerig, "Unbiased diffeomorphic atlas construction for computational anatomy." *Neuroimage*, vol. 23 Suppl 1, pp. S151–S160, 2004. ˆ174

[64] K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, *Statistical parametric mapping: the analysis of functional brain images*. Academic Press, Elsevier, London, 2007. ˆ175

[65] C. H. Salmond, J. Ashburner, F. Vargha-Khadem, A. Connelly, D. G. Gadian, and K. J. Friston, "The precision of anatomical normalization in the medial temporal lobe using spatial basis functions." *Neuroimage*, vol. 17, no. 1, pp. 507–512, Sep. 2002. ˆ175

[66] J. R. Chumbley and K. J. Friston, "False discovery rate revisited: Fdr and topological inference using gaussian random fields." *Neuroimage*, vol. 44, no. 1, pp. 62–70, Jan. 2009. ˆ176

[67] J. T. Devlin and R. A. Poldrack, "In praise of tedious anatomy." *Neuroimage*, vol. 37, no. 4, pp. 1033–41; discussion 1050–8, Oct. 2007. ˆ176

[68] K. J. Friston, K. E. Stephan, T. E. Lund, A. Morcom, and S. Kiebel, "Mixed-effects and fMRI studies." *Neuroimage*, vol. 24, no. 1, pp. 244–252, Jan. 2005. ˆ177

[69] Y. Hochberg and A. Tamhane, *Multiple comparison procedures*. John Wiley & Sons, 1987. ˆ177

[70] J. Ludbrook, "On making multiple comparisons in clinical and experimental pharmacology and physiology." *Clin Exp Pharmacol Physiol*, vol. 18, no. 6, pp. 379–392, Jun. 1991. ˆ177

[71] K. J. Friston, "Testing for anatomically specified regional effects." *Hum Brain Mapp*, vol. 5, no. 2, pp. 133–136, 1997. ˆ177

[72] F. L. Bookstein, ""Voxel-based morphometry" should not be used with imperfectly registered images." *Neuroimage*, vol. 14, no. 6, pp. 1454–1462, Dec. 2001. ˆ177

[73] K. Amunts, A. Schleicher, and K. Zilles, "Cytoarchitecture of the cerebral cortex–more than localization." *Neuroimage*, vol. 37, no. 4, pp. 1061–5; discussion 1066–8, Oct. 2007. ˆ177

[74] D. K. Jones, M. R. Symms, M. Cercignani, and R. J. Howard, "The effect of filter size on VBM analyses of DT-MRI data." *Neuroimage*, vol. 26, no. 2, pp. 546–554, Jun. 2005. ˆ178

[75] W. R. Crum, L. D. Griffin, D. L. G. Hill, and D. J. Hawkes, "Zen and the art of medical image registration: correspondence, homology, and quality." *Neuroimage*, vol. 20, no. 3, pp. 1425–1437, Nov. 2003. ˆ178

Chapter 4

Multivariate Morphometry

The purpose of this chapter is to provide a thorough theoretical and practical study of the major extensions of mass-univariate statistical analyses like voxel-based morphometry to voxel-wise 'mass-multivariate' analyses.

The theory of multivariate or generalised tensor-based morphometry is expounded, and an attempt is made to provide a novel synthesis of two complementary viewpoints: a physical solid-mechanics approach, with the goal of aiding intuition; and a mathematical group-theoretic approach, which provides rigour.

Issues related to spatial normalisation of longitudinal deformation fields, Jacobians or strain tensors are thoroughly dealt with, in a discussion which also contributes to the closely related problem of diffusion tensor reorientation.

The permutation-testing framework described in chapter 2 is employed to furnish inferences corrected for the multiple testing problem, and to allow the use of two test statistics which have never before been applied to morphometric data in this way. The second such statistic permits the analysis of the principal direction of strain; one of a number of quantities compared and contrasted to other strain-tensor-derived measures. Also included, are analyses of the displacement fields, and multivariate voxel-wise analyses of both displacement and volume changes using a local 'searchlight' kernel [1].

This chapter can be seen as an attempt to develop the equivalent 'Morphometry' chapter from Ashburner's thesis [2, Ch. 6]; the main progress with respect to that work (which focussed strongly on a solid-mechanics interpretation of strain tensors) stems from broader consideration of related fields, for example bringing to morphometry practical advances from diffusion tensor imaging and the theoretical developments in Riemannian tensor metrics.

4.1 Introduction

Shape is fundamentally a multivariate concept, and one of the strongest criticisms of Voxel-Based Morphometry is its univariate limitation [3]. As illustrated by Mechelli et al. [4], the use of total tissue volume as a covariate can partially address this concern, allowing questions about the local volume after adjusting for global effects. However, the nature of shape suggests that multivariate analysis may offer significant potential to further the

understanding of anatomical differences between subjects and of morphometric changes over time.

Fully multivariate modelling of three-dimensional images (see e.g. [5, 6, 7]) requires significant data reduction and/or regularisation in order to cope with the dimensionality of the data vastly exceeding the number of scans. Furthermore, by considering entire images as observations, these techniques complicate the anatomical interpretation of their results. In this chapter, we investigate methods that lie between the two extremes of the mass-univariate and the massively high-dimensional. Linear statistical modelling is employed, with familiar concepts of design matrices and contrasts, producing standard thresholded maps of significant voxels. However, the model is applied to data which has a multivariate observation at each voxel. Such observations can either come from measurements which are inherently multivariate (for example the three-dimensional displacement at each voxel, from a non-rigid registration), or from collecting together the univariate observations of all voxels within a specified neighborhood of the current voxel into a single multivariate summary — Kriegeskorte et al.'s 'searchlight' [1].

The theory and implementation developed in chapter 2 and appendix D provide the necessary tools for efficient permutation-based linear modelling of these data. Below, we will describe various sources of multivariate measurements with potential application to morphometry, focussing in particular on those derived from the Jacobian matrix. Theoretical concepts are explored, including the issue of inter-subject normalisation of longitudinal information. These approaches will then be investigated in practice through analyses of the DRC/GSK MIRIAD longitudinal Alzheimer's Disease data-set [8]. The SPM software is used to perform the non-rigid registrations from which we derive the multivariate measures and the univariate data suitable for analysis with the searchlight technique.

4.1.1 Summary of potential applications

The following is a concise overview of some potential applications of the multivariate statistical methods mentioned above to various sources of data, within the field of longitudinal MR imaging of dementia.

Serial data

The most obvious potential source of multivariate voxel-wise data is to consider the multiple measurements of the longitudinal time-series as a single multivariate measurement at each voxel. This is a standard approach for the analysis of repeated measures models, which is an alternative to univariate ANOVA with non-sphericity correction [9]. It is common practice not to analyse the original m time-point multivariate measurements, but rather the data after transformation by some contrast. For example, the $m-1$ differences with respect to the first time-point, or the $m-1$ adjacent differences $(t_2 - t_1, t_3 - t_2, ..., t_n - t_{n-1})$. Polynomial contrasts using linear, quadratic, etc. terms can also be used. Exact representation of m time-points requires m polynomial terms from constant (degree 0) through to $(m-1)$th degree, though, as with the differencing contrasts, it seems common to reduce the dimensionality to $m-1$, in this case by neglecting

the average (constant) term. The theory of the multivariate general linear model (A.4) seems to apply just as well to the full m-variate data, but the use of contrasts simplifies the interpretation of results. A significant multivariate finding can be directly followed by 'protected' univariate tests of the particular differences or polynomial terms, within the framework of Fisher's 'least significant difference' multiple comparison procedure (see section 1.6.4).

By combining time-points in this way, there is an implicit assumption that the different times are equivalent across subjects. Clearly, this framework cannot accommodate unbalanced designs. Deliberately unbalanced experiments might seem desirable in some cases, for example measuring patients on more occasions than controls, since the patients might be expected to be changing more rapidly. More commonly, initially balanced designs can suffer from missing data, for example patients becoming too ill to continue in the study. Equivalence of time-points also means that there must not be significant variation in acquisition times for the different subjects' measurements. Mixed models, discussed in section 1.6.3, can avoid some of these limitations, but the added computational complexity seems likely to preclude the combination of variance-component estimation with permutation testing.

Non-standard or multi-spectral MRI data

In addition to segmented tissue-density, which forms the basis of voxel-based morphometry, complementary information can be derived from other sources, including different MRI sequences and other imaging modalities. For example, Hayasaka et al. [10] analysed tissue density and perfusion-weighted images from arterial spin labelling (ASL) MRI, finding both areas of concordance and discordance in the two measures. Analysis of dementia could also benefit by including information from

- *quantitative* images of T1, T2 or other parameters [11]
- diffusion-weighted information, to investigate markers of axonal degeneration
- PET imaging, e.g. using PIB [12] to measure the distribution of amyloid deposits
- FLAIR MRI, to investigate white-matter lesions and differentiate vascular dementia

These, and other imaging modalities were mentioned briefly in sections 1.1 and 1.3.2.

Such combined data-sets should meet the assumptions of equivalence over subjects more easily than the serial data discussed above. However, it may still be difficult to model or interpret large differences in scale, variability, etc. between the data sources. Furthermore, with such disparate data, more complex relationships could be of interest, including dissociation. For these reasons, the combining function framework used by Hayasaka et al. [10] (and discussed in section 2.3.4) might be superior for such applications.

Deformation-based morphometry

Once non-rigid deformations have been found that register multiple images, the resultant displacement fields contain information on the estimated correspondences. We will use

the term deformation-based morphometry (DBM) to refer generally to approaches using this concept. It is helpful to distinguish between high-dimensional 'image-wise' analysis of complete deformation fields, and low-dimensional voxel-wise multivariate analysis of displacement vectors. Ashburner et al. [13] pioneered the former approach; Gaser et al. [14] focussed on the latter technique, in work more similar to that which is presented here.

With serial data, the deformation fields from within-subject registration can be analysed, using spatial normalisation simply to achieve intersubject correspondence (see section 4.2.10 below). Cardenas et al. [15] have used this technique to look at brain changes caused by chronic alcoholism.

Diffusion tensor MRI

In the context of Diffusion Tensor Imaging, Whitcher et al. [16] discuss the common practice of reducing the tensor to a scalar measure, such as Fractional Anisotropy, in order to perform voxel-wise statistical testing; they point out that this involves a loss of information, and suggest that multivariate testing should offer several advantages. Below we describe a family of symmetric positive definite strain tensors derived from the gradient of a non-rigid spatial transformation. There is a close analogy between diffusion and strain tensors, which means that many of the approaches applied to the former (including Whitcher et al.'s multivariate analysis) may be usefully transferred to morphometry. Whitcher et al. investigate three alternative multivariate tests: the two-sample Hotelling's T^2 (a standard parametric test) [17]; the Cramér test (as presented in section 4.3.3); and a test based on non-parametric combination of dependent multivariate permutation tests (using the Fisher combining function) [18]. In [16], multiple tests are dealt with using false discovery rate (FDR) correction, in contrast to the family-wise error (FWE) controlling permutation-testing methods used here.

Tensor-based morphometry

The spatial derivative of a three-dimensional deformation field is a three-by-three matrix, or second order tensor, known as the Jacobian.[1] The term tensor-based morphometry (TBM) is usually taken to imply the analysis of a measure derived from the Jacobian. In TBM, as in DTI above, it is also common practice to reduce the information in the Jacobian tensor to a single scalar (or zeroth order tensor) given by its determinant at each voxel [19, 20]. Ashburner and Friston [21] suggested that a better approach might be multivariate analysis using one of several possible Lagrangian strain tensors derived from the polar decomposition of the Jacobian matrix. The theory underlying this is expanded upon below.

Both cross-sectional and longitudinal transformations have been analysed using multivariate TBM. Lepore et al. [22, 23] apply their method to a cross-sectional comparison of 26 HIV/AIDS patients compared with 14 matched healthy controls. Studholme and

[1] Some authors use the term Jacobian to refer to the determinant of this matrix, explicitly referring to the 'Jacobian matrix'. Both conventions are common; we will typically consider the Jacobian to be the matrix, and will explicitly refer to the Jacobian determinant.

Cardenas [24] study the brain changes in recovering alcoholics, on 16 consistent abstainers and 8 relapsers (no non-alcoholic controls), over two time-points approximately 8 months apart. When longitudinal change is the focus, spatial normalisation is still required to define inter-subject correspondence. In section 4.2.10 we will discuss the issue of how the inter-subject transformation interacts with the intra-subject one.

The multivariate searchlight

The basic premise of the work of Kriegeskorte et al. [1, 25] is to look for information in the multivariate pattern of the observed voxels in a local window around each voxel; this multivariate searchlight is then scanned through the image. By avoiding the preprocessing step of smoothing,[2] searchlight has the potential to detect more complex patterns, whose structure would be damaged by naïve spatial blurring. As a specific example, TBM of AD may fail to find (or fail to precisely locate) atrophy in a small structure like the hippocampus, if the shrinking structure is adjacent to expanding CSF, and these opposite effects are averaged together via the smoothing kernel. We explore the searchlight in greater detail below.

4.1.2 Statistical methods applied to multivariate morphometry

We now provide a brief, but, to the best of our knowledge, complete survey of the different types of statistical analysis applied to multivariate tensor-based morphometry. Univariate TBM has typically used either the SPM implementation of the parametric GLM or a simple form of permutation testing, though we do not attempt to survey in detail the much larger number of publications employing univariate analysis.

The work of Gaser et al. [14, 26] on deformation-based morphometry employed uncorrected statistics, though they alluded to the availability of random field theory (RFT) results for Hotelling T^2 fields, and to the possibility of approximate transformation from Wilks' Λ to Snedecor's F, as discussed in section A.4.4, which is also used by Studholme and Cardenas [24]. Cao and Worsley [27] tested the three components of inter-subject deformation fields, using RFT results for Hotelling T^2.

Ashburner [2] noted at the time of writing that RFT results were not available for a general Wilks' Λ field, and he presented only uncorrected statistics. RFT results for Wilks' Λ are now available from Carbonell et al. [28],[3] but they appear not to be widely adopted.

In order to apply RFT results to Hotelling's T^2 or Wilks' Λ, it is necessary to have an estimate of the smoothness of the multivariate residuals. Interestingly, Worsley's original work included this multivariate generality [29], but it seems to have been lost from a more recent development [30] which formed the basis for the implementation in the popular SPM software.

[2] Note that even with permutation testing, which removes the requirement for normally distributed data, smoothing (or the searchlight) is typically still beneficial, either to compensate for residual mis-registration, or to enhance the signal-to-noise ratio, or perhaps to reduce the number of very small scattered findings.

[3] See also `http://www.math.mcgill.ca/keith/felix/felix.htm`

Ashburner [2] found some evidence of non-normality in all the strain tensors he considered, which motivates the use of non-parametric testing.

Lepore et al. use non-parametric permutation testing, based on the squared Mahalanobis distance [22] or permutationally equivalent Hotelling T^2 statistic [23], with 5000 permutations. They produce uncorrected p-values at each voxel, before assessing the overall significance of the p-map using Storey's 'positive False Discovery Rate' [31, 32]. Their visualisation, however, focusses on uncorrected p-values.

Studholme and Cardenas [24] (discussed further below) use a special case of Wilks' Λ for a single interest covariate, but report only raw statistic maps thresholded at an arbitrary level (transformed F=2). Interestingly, their 24 subjects seems not to be sufficient for estimating the covariance matrices, which, for their multivariate data (m=9) will have 45 unique elements. There appears to be no discussion of the seemingly necessary shrinkage or regularisation techniques (see section 4.3.3).

Statistical contribution of the present work

To the best of our knowledge, this chapter represents the first application of FWE-controlling multivariate permutation testing to deformation- or tensor-based morphometry (or to the searchlight). It is also believed to include the first use of the two-sample Cramér statistic with morphometric data, and the first use of the bipolar Watson statistic with orientational information in the context of morphometry. This work is also believed to feature the first neuroimaging application of FWE-controlling permutation-testing to either the Cramér or Watson statistics. Additionally, this seems to be the first time a step-down FWE-controlling permutation method has been applied in structural neuroimaging.

4.2 Theory

4.2.1 The searchlight

Kriegeskorte et al. [1, 25] analyse the multivariate observations formed from accumulating the univariate voxel-wise data within a discretised approximation to a spherical kernel around each voxel (i.e. a spherical searchlight is swept through the image). They explore a range of kernel volumes from 1 to 123 voxels (for the properties of these see table 4.3).

Kriegeskorte et al. analyse the Mahalanobis distance between two fMRI conditions. The squared Mahalanobis distance is very closely related (and permutationally equivalent) to Hotelling's two-sample T^2 test, which is in turn a special case of the more general likelihood-ratio based Wilks' Λ statistic. Therefore the multivariate permutation testing methods proposed here in chapter 2 and appendix D generalise Kriegeskorte et al.'s method to a wide range of linear models. Thus far, the searchlight method has been applied using permutation-based testing to generate uncorrected p-values, which are then corrected using the False Discovery Rate mechanism [33]. Therefore another development of this chapter is the presentation of the first FWE-corrected searchlight analyses.

Because spatial smoothing is essentially an averaging process, it increases the signal-to-noise ratio (SNR) of spatially distributed 'blob-like' signals. The matched filter theorem

[34] implies that the analysis will be most sensitive to signals matching the shape and size of the smoothing kernel. By replacing the explicit averaging with a more general combination of signals via multivariate statistics, the searchlight hopes to achieve almost the same SNR without assuming such a simple form for the signal. However, the searchlight does not avoid the problem that different scales of signals will be best identified with different size kernels. In fact, the problem is greatly exacerbated, since larger scale patterns will require large searchlight kernels, leading to multivariate observations of undesirably high-dimensionality with respect to the number of images.

We argue that patterns with a very large spatial extent are less likely simultaneously to exhibit fine spatial detail. This motivates the application of multi-resolution or scale-space techniques; in particular, pyramid approaches [35], which effect the necessary reduction in dimensionality as they decrease the high-resolution content.

Multi-resolution methods

Multi-resolution techniques are common in image-processing, particularly in image registration, where they can be expected to improve accuracy, robustness [36], and speed, by ensuring that (fast) initial computations at lower resolutions bring the coarser image features into rough alignment [37], helping to avoid local extrema in the objective function, and reducing the number of expensive full-resolution iterations required [38].

Pyramid methods are also useful in analyses, particularly fully multivariate ones such as image-classification using Support Vector Machines, where they can provide a helpful dimensionality reduction in addition to matching the scale of expected patterns. Such techniques range from simple down-sampling via voxel-averaging [39], through Gaussian blurring and down-sampling, to more complex techniques such as those found in the work of Davatzikos' group [40, 41].

The simplest voxel-averaging techniques are likely to have poor frequency-domain properties, particularly if upsampling is also required. Unser et al. [42] proposed the use of spline-pyramids derived for optimal least-squares representation for a given approximation order. They note that 'among all interpolants of a given degree of smoothness, [polynomial splines] are those that oscillate the least' and go on to show superior performance compared to linear and Gaussian pyramids. In later work, Brigger et al. [43] extended these spline pyramids to symmetric centred-topology versions with several advantages, including 'more faithful image representation at coarser pyramid levels' [43] which is particularly appealing for the present purpose. The techniques also offer an optimal means to upsample after down-sampling (e.g. to return statistical results at the lower resolution to the full structural template resolution for visualisation) which is helpful here. Using C code made available by the above-cited authors (`http://bigwww.epfl.ch/sage/pyramids/`), we have incorporated this approach into our MATLAB-based permutation-testing software.

Wavelet methods [40, 44, 45] probably provide the most sophisticated way of performing multi-resolution analysis, without the naïvety of stationary and isotropic Gaussian smoothing or spline-pyramid downsampling. Wink and Roerdink [46] perform a direct

comparison of Gaussian smoothing and Wavelet denoising in the context of statistical parametric mapping, strongly favouring the latter approach. However, due to the large number of choices to be made regarding mother-wavelet and filter-orders, etc. and the greater complexity in implementation, we postpone further consideration of this for future work.

4.2.2 Displacement and deformation vector fields

Deformation-based morphometry is taken here to mean voxel-wise analysis of displacement vector fields derived from non-rigid registration, either between subjects, or over time after inter-subject spatial normalisation. A point $r_0 = (x_0, y_0, z_0)$ in the target image (defined over a domain Ω_0) is mapped to a corresponding point $r_1 \in \Omega_1$ in the source image by a transformation T_{10}.[4] Distinction should be drawn between the transformation $r_1 = T_{10}(r_0)$, which we consider as a voxel-wise 'deformation field' connecting points r_0 and r_1, and the 'displacement field' $u_{10} = (u^x_{10}, u^y_{10}, u^z_{10})$ which measures the offset from an identity transformation, $r_1 = r_0 + u_{10}(r_0)$. The latter are often associated with small-deformation registration approaches (e.g. using an elastic penalty on the displacement away from an identity) in contrast to large deformation approaches, which focus on the transformation (e.g. penalising instead a velocity field from which the transformation is derived). However, for any transformation, we may compute $u_{10}(r_0) = T_{10}(r_0) - r_0$. For the purpose of standard statistical analysis of DBM data, however, we may note that the identity transformation is common to all subjects, and therefore both deformation and displacement fields will yield the same results for statistical models where the constant term is in the space of the nuisance covariates. This is also true after smoothing, thanks to linearity, though care is required that image boundaries are handled properly (or are outside the analysis mask), since zero-padding is incorrect for deformation fields. Similarly, the spline-pyramid downsampling discussed in 4.2.1 assumes reflectant boundary conditions which would not be appropriate for deformation fields.

4.2.3 The Jacobian tensor field

The Jacobian of a deformation vector field is a tensor field. At each point, the Jacobian matrix (of partial derivatives) relates infinitesimal vector elements in the target and source:

$$\begin{pmatrix} dx_1 \\ dy_1 \\ dz_1 \end{pmatrix} = \begin{pmatrix} \frac{\partial x_1}{\partial x_0} & \frac{\partial x_1}{\partial y_0} & \frac{\partial x_1}{\partial z_0} \\ \frac{\partial y_1}{\partial x_0} & \frac{\partial y_1}{\partial y_0} & \frac{\partial y_1}{\partial z_0} \\ \frac{\partial z_1}{\partial x_0} & \frac{\partial z_1}{\partial y_0} & \frac{\partial z_1}{\partial z_0} \end{pmatrix} \begin{pmatrix} dx_0 \\ dy_0 \\ dz_0 \end{pmatrix}$$

$$dr_1 = \frac{\partial r_1}{\partial r_0} dr_0,$$

[4] Note that r and s etc. denote vectors, e.g. $r_0 = [x_0\ y_0\ z_0]^T$, but there is no need to distinguish them from scalars here. Arguably, transformations should be written with lower-case, since $T(r)$ is a vector field, but the notation used here seems less likely to cause confusion when T is mentioned in isolation.

or, emphasising that the Jacobian is defined at each point r_0

$$\begin{aligned} dr_1 &= \left.\frac{\partial T_{10}(r)}{\partial r}\right|_{r_0} dr_0 \\ &= J_{10}(r_0)\, dr_0. \end{aligned}$$

The Jacobian is also known as the 'deformation gradient tensor', and is related to the 'displacement gradient tensor' [47], $K = \frac{\partial u}{\partial r} = J - I$,[5] consistent with our distinction between displacement and deformation vector fields.

The absolute value of the Jacobian determinant appears in the expression for a change of variables in a multivariate integral:

$$\int_{\Omega_1} f(r_1)\, dr_1 = \int_{\Omega_0} f(T_{10}(r_0)) \left|\det\left(\frac{\partial r_1}{\partial r_0}\right)\right| dr_0,$$

where $\Omega_0 = T_{10}^{-1}(\Omega_1)$.[6] A special case relates total volumes (e.g. of segmented structures) in the source and target spaces [49]:

$$\int_{\Omega_1} dr_1 = \int_{\Omega_0} |J|\, dr_0.$$

More intuitively, an infinitesimal cube maps to a parallelepiped, whose volume is given by the original cube's volume multiplied by the determinant of the Jacobian. I.e. $|J|$ indicates the local volume change due to the transformation; $|J| = 0$ implies that one or more dimensions have been flattened (e.g. a volume has been transformed to a plane, line or point), while a negative determinant indicates 'folding' or 'tearing' of space has occurred. For an affine transformation $T(r) = Lr + b$, the Jacobian is simply given by the linear part, $\partial(Lr + b)/\partial r = L$, which is constant. Note also that $|J| = |L| = |A|$ where A is the homogeneous form of the transformation matrix (a 4×4 matrix with the linear part in the upper-left block, b to its right, and a final row of [0 0 0 1]). Rotations preserve volume and have $|L| = 1$ as expected, while reflections have $|L| = -1$.

4.2.4 Unified deformation-based morphometry

DBM and a variant of TBM can be placed within a unified statistical framework of methods derived from the deformation field [50]. In particular, Chung et al. [50] note that the 'volume dilatation' given by the trace of the displacement gradient tensor ($\mathrm{tr}\,(K) = \partial u_x/\partial x + \partial u_y/\partial y + \partial u_z/\partial z$) is statistically independent of the displacement vector field components. This eases the interpretation of multiple statistical tests, however, it should be noted that the dilatation is only an approximation to the volume change given by the

[5] Subscripts will be omitted from T, J, etc. if they are obvious from the context or are not of interest.

[6] In practice, rigid or affine transformations are often used initially to align the images approximately, after which, for simplicity, the non-rigid transformations are often assumed to have fixed boundaries [48], meaning the range and domain are the same: $\Omega_1 = \Omega_0 = \Omega$.

determinant of the transformation's Jacobian matrix, as we now show.[7]

$$
\begin{aligned}
r_1 &= T_{10}(r_0) = r_0 + u_{10}(r_0) \\
J_{10} &= \frac{\partial r_1}{\partial r_0} = \frac{\partial r_0}{\partial r_0} + \left.\frac{\partial u_{10}}{\partial r}\right|_{r_0} = I + K_{10}(r_0) \\
|J| &= |I + K| \qquad \text{(dropping subscripts)} \\
&= \prod_{i=1}^{3} \lambda_i(I + K) \\
&= \prod_{i=1}^{3} (1 + \lambda_i(K)) \\
&= 1 + \lambda_1(K) + \lambda_2(K) + \lambda_3(K) \\
&\quad + \lambda_1(K)\lambda_2(K) + \lambda_2(K)\lambda_3(K) + \lambda_3(K)\lambda_1(K) \\
&\quad + \lambda_1(K)\lambda_2(K)\lambda_3(K) \\
&\approx 1 + \lambda_1(K) + \lambda_2(K) + \lambda_3(K) = 1 + \mathrm{tr}\,(K) = \mathrm{tr}\,(J) - 2.
\end{aligned}
$$

This suggests that the determinant should be prefered for larger deformations; we investigate the practical impact of the difference on our data below. An alternative interpretation of the dilatation comes from the fact that it is the divergence of the displacement vector field. The divergence theorem [51] implies that the volume integral of the divergence over a particular region is equal to the surface integral of the flux through the region's boundary:

$$
\iiint_{\Omega} (\nabla \cdot u)\, d\omega = \iint_{\partial\Omega} u \cdot \hat{n}\, dS, = \iint_{\partial\Omega} u \cdot dn, \tag{4.1}
$$

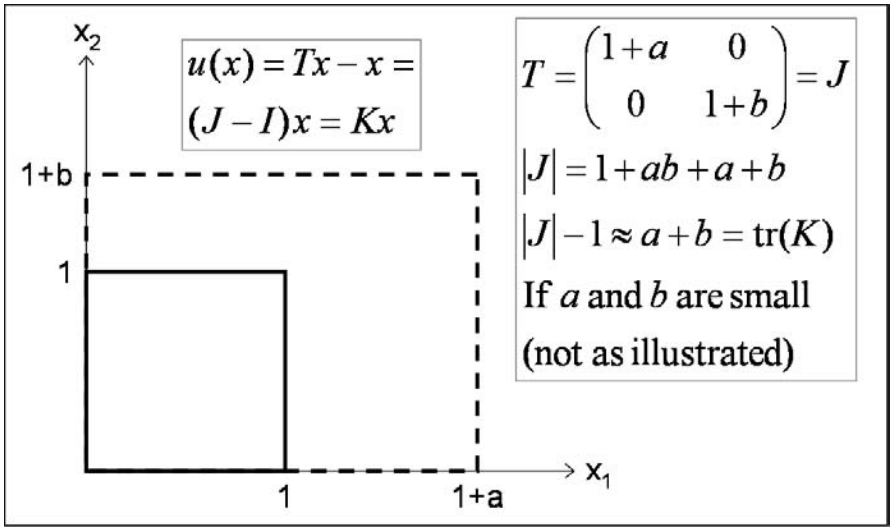
where $\hat{n}$ is the outward-pointing unit vector normal to the boundary $\partial\Omega$ of the volume Ω.

In the case of a displacement field, the surface integral of the 'flow' out of the boundary appears intuitively equivalent to the volume increase of the region described by that displacement field. It is initially difficult to tally this with the fact that the integral of the Jacobian determinant over a region also gives that region's transformed volume. The answer, as suggested above, is that the outward flow only gives the increase in volume for small deformations — for larger changes, the change in the boundary over which the flow is considered must be taken into account. Figure 4.1 illustrates this for a simple example in two dimensions (with exaggerated deformation).

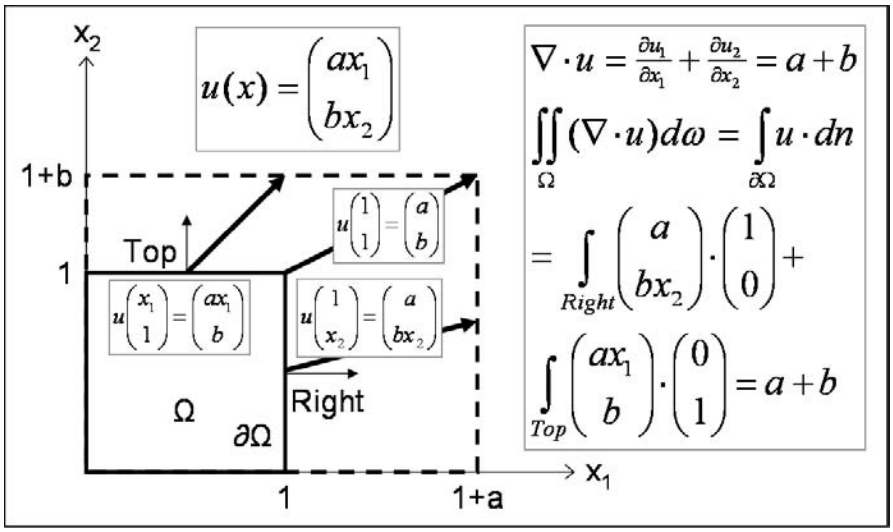
4.2.5 Strain tensors

In this section, we review material from several sources relating to the study of shape via tensors derived from the Jacobian or deformation gradient tensor. First, the perspective is from the field of solid mechanics; later, a more abstract mathematical viewpoint will be considered.

[7] We use λ_i here as an operator that returns the eigenvalues, such that $\lambda_i(A)$ and $\lambda_i(B)$ are the i^{th} eigenvalues of A and B respectively.



(a) Transformation and Jacobian



(b) Displacement and divergence

Figure 4.1: Comparison of volume change and volume dilatation, for transformation of the unit square in 2D. (a) For a simple linear transformation — here chosen as anisotropic scaling — the Jacobian matrix is equal to the transformation matrix. The volume (area) gained is $|J| - 1$ which can be approximated as $a + b$ if a and b are small enough to neglect their product (which is *not* the case in this illustration). (b) Considering instead the divergence of the displacement field, the (1D) surface integral of the flow through the faces of the unit square is also given by the (2D) volume integral of the dilatation, thanks to the divergence theorem. Note though that the flow out of a hypothetical rectangle intermediate between the initial and final result in fact has a larger boundary. Hence Chung et al.'s dilatation [50] is only truly appropriate for small deformations.

Since the Jacobian relates infinitesimal vector elements, it contains information about the local change in length and orientation of such elements, which, in general, depends on the orientation of the original element. Equivalently, one may interpret the Jacobian tensor in terms of local strains, shearing and reorientation. Shears can be seen more simply as strains along rotated axes, meaning that the Jacobian matrix can be decomposed solely into strains and rotations. This can be achieved via the Singular Value Decomposition (A.2), which gives $J = XSZ^T$.[8] Because the Jacobian is a square 3×3 matrix, each of the components in the decomposition are also 3×3 matrices. X and Z have orthonormal columns and hence satisfy $X^T = X^{-1}$ in common with rotation (or reflection) matrices. If we assume there is no folding and no singularities in the deformation $|J| = |X| \cdot |S| \cdot |Z| > 0$, meaning both these orthogonal matrices can be chosen to have positive determinants and hence can be interpreted as rotations. S is a diagonal matrix of strictly positive values, which can be interpreted as scalings along three orthogonal (rotated) axes.

From the singular value decomposition, one may derive two versions of the polar decomposition:

$$\begin{aligned} J &= XSZ^T \\ &= X(Z^TZ)SZ^T = (XZ^T)(ZSZ^T) = RU \\ &= XS(X^TX)Z^T = (XSX^T)(XZ^T) = VR. \end{aligned}$$

$R = XZ^T$ is the product of two rotations and hence is a rotation. U and V are symmetric positive definite (for non-singular J) tensors known as the right and left stretch tensors [47]. They can be related to the Jacobian via $J^TJ = ZS^2Z^T$ and $JJ^T = XS^2X^T$, since appendix B.1.1 then yields $V = \left(JJ^T\right)^{1/2}$ and $U = \left(J^TJ\right)^{1/2}$. Alternatively, $U^2 = J^TJ = C$, where C is sometimes known as the Cauchy-Green deformation tensor [52]. Lepore et al. [23] refer to U somewhat non-standardly as simply the 'deformation tensor'.

Strain tensors that can be derived from U are known as Lagrangian strain tensors, while tensors derived from V are known as Eulerian. Expanding upon an explanation given by Ashburner and Friston [21], the Lagrangian frame is appropriate for computational anatomy, since multiple Jacobians, $J^{(n)} = R^{(n)}U^{(n)}$, can be analysed in terms of their right stretch tensors $U^{(n)}$ in the template space, ignoring their differing (post-) rotations to different source image spaces.

As discussed by Ashburner and Friston [21], each frame of reference possesses an entire family of different strain tensors with varying properties, summarised here in table 4.1. The Biot tensor relates to the solid-mechanics concept of 'nominal strain', while the Hencky tensor (which we return to below) is related to the concept of logarithmic strain [47].

The strain tensor G is one of the most commonly used in solid-mechanics, but it is unfortunately also one of the least consistently named. It is variously known as the 'finite strain tensor', 'Lagrange strain tensor' [47], 'Lagrangian strain tensor' [53] 'Green

[8] We write the SVD with the matrices of left and right singular vectors as X and Z instead of U and V as elsewhere in this thesis, allowing the latter to be used for the right and left stretch tensors respectively, as is conventional in continuum mechanics [47].

Name	m	$E_m = \frac{U^m - I}{m}$
Almansi	-2	$\frac{I-U^{-2}}{2} = \frac{I-C^{-1}}{2}$
Hencky	$\lim_{(m\to 0)}$	$H = \mathrm{logm}\,(U)$
Biot	1	$U - I$
Green	2	$G = \frac{U^2-I}{2} = \frac{C-I}{2}$

Table 4.1: Lagrangian strain tensors E_m derived from the right stretch tensor $U = (C)^{1/2} = \sqrt{J^T J} = ZSZ^T$ where $J = XSZ^T$.

(Lagrangean[9]) strain tensor' [21] or 'Green - Saint-Venant' tensor [52].[10] Note that G is closely related to the Cauchy-Green deformation tensor $C = J^T J$, and that C may also be expressed in terms of the displacement gradient tensor K, giving

$$C = J^T J = (I + K)^T (I + K) = I + K^T + K + K^T K \tag{4.2}$$

$$G = \frac{J^T J - I}{2} = \frac{C - I}{2} = \frac{K^T + K + K^T K}{2}. \tag{4.3}$$

For isotropic materials satisfying Hooke's law, the Saint-Venant - Kirchoff elasticity energy is related to the finite strain tensor by

$$\int_\Omega \mu \,\mathrm{tr}\left(G^2\right) + \frac{\lambda}{2}\mathrm{tr}^2\left(G\right) dr,$$

where μ and λ are the Lamé coefficients [52]. This energy can be used to derive the constitutive equations; further details on its mathematical properties are discussed in [54]. It is interesting to note that if $\lambda = 0$ the energy reduces to

$$\begin{aligned}
\mu \int \mathrm{tr}\left(G^2\right) &= \frac{\mu}{4} \int \mathrm{tr}\left((C-I)^2\right) \\
&= \frac{\mu}{4} \int \|C - I\|_F^2 \\
&= \frac{\mu}{4} \int d^2_{\mathrm{Euc}}(C, I),
\end{aligned}$$

which is the squared Euclidean distance between the Cauchy-Green deformation tensor $C = J^T J$ and the identity. Pennec et al. [52] consider replacing this Euclidean metric with a Riemannian one that accounts for the curvature of the space of symmetric positive definite tensors, leading to their concept of 'Riemannian Elasticity'. Pennec et al. are concerned with nonlinear elastic regularisation of deformation fields, but their work is also closely related to Riemannian analysis of tensors, as discussed in section 4.2.6.

For small deformations, the term $K^T K$ in equations 4.2 and 4.3 can be ignored, leading to the infinitesimal strain tensor $F = \frac{K^T + K}{2}$ (which also approximates the Eulerian finite strain tensor, thanks to the small difference between the coordinate frames) [47]. This

[9] The Oxford English Dictionary gives Lagrangian as the main spelling; Lagrangean as an alternative.

[10] The final expression is also sometimes misleadingly typeset, e.g. as 'Green-Saint Venant', but its name derives from George Green and Adhémar Jean Claude Barré de Saint-Venant.

tensor is linear in the displacement, which simplifies some analyses, but is only suitable for small strain situations. Volumetric infinitesimal strain [47] is defined as $\mathrm{tr}\,(F) = \mathrm{tr}\,(K) = \nabla \cdot u$, which is equivalent to Chung's dilatation [50] again reflecting our earlier argument that this is less suitable for larger deformations. Bower also defines a 'deviatoric' infinitesimal strain tensor as $F - I\frac{\mathrm{tr}(F)}{3}$ (closely related to fractional anisotropy in DTI) and an infinitesimal rotation tensor, given by the skew symmetric matrix $\frac{K-K^T}{2}$. Both of these concepts are returned to in section 4.2.9.

The Hencky tensor is the matrix logarithm of U. The matrix exponential and logarithm are defined and explored in appendix B. The singular and eigenvalue decompositions of U are respectively $U = ZSZ^T = ZSZ^{-1}$; the latter allows the use of (B.8) to show that the Hencky tensor derives from U simply by taking the scalar logarithm of the latter's eigenvalues. Using (B.11), the Hencky tensor can also be expressed as

$$H = \mathrm{logm}\,(U) = \mathrm{logm}\left(\left(J^T J\right)^{1/2}\right) = \tfrac{1}{2}\mathrm{logm}\left(J^T J\right). \tag{4.4}$$

showing that the (real) eigenvalues of the symmetric Hencky tensor are equal to the logs of the (positive) singular values of the Jacobian tensor, or equivalently, the logs of the square roots (or halves of the logs) of the eigenvalues of the symmetric positive definite Cauchy-Green deformation tensor $C = J^T J$. The trace of the Hencky tensor is equal to the sum of its eigenvalues, which is the log of the product of eigenvalues of $U = \left(J^T J\right)^{1/2}$, and hence is equal to the commonly analysed log of the determinant of the Jacobian:

$$\mathrm{tr}\,(H) = \mathrm{tr}\left(\mathrm{logm}\left(\left(J^T J\right)^{1/2}\right)\right) = \log|\left(J^T J\right)^{1/2}| = \log|J|. \tag{4.5}$$

Ashburner et al. [55] use the fact that the squared Frobenius norm of the Hencky tensor is given by the sum of the squares of the logs of the singular values of the Jacobian to motivate priors for a Bayesian regularisation of registration:

$$\begin{aligned}\|H\|_F^2 = \sum_{i,j} H_{ij}^2 &= \mathrm{tr}\left(H^T H\right) = \mathrm{tr}\left(V\,\mathrm{logm}^2\,(S)\,V^T\right) \\ &= \mathrm{tr}\left(\mathrm{logm}^2\,(S)\right) = \sum_i \log^2(s_i).\end{aligned}$$

The logs of the singular values have two useful properties in this respect, as explained by Ashburner: (i) any probability distribution over $\log s_i$ prevents meaningless 'negative' lengths, unlike a prior on e.g. the eigenvalues of U, which would need to constrain them to be positive, (ii) if the $\log s_i$ — related to lengths — are assumed to be normally distributed, then so are their sums $s_1 + s_2$ and $s_1 + s_2 + s_3 = \log|J|$ — which relate to areas and volumes respectively. The experimental results in this chapter use the same (high-dimensional warping) registration algorithm developed by Ashburner et al. [55], motivating a focus on the Hencky tensor (and measures derived thereof) in our analysis.

4.2.6 Vector spaces, groups and manifolds

This section provides a more abstract mathematical perspective on the analysis of (Jacobian) matrices. It builds gradually from simple cases (the full relevance of which will become apparent only by contrast to the later examples) culminating in an alternative motivation for analysing the Hencky tensor. While there is no novel theoretical development in this section, it is hoped that it provides a more approachable introduction, and perhaps a more unified synthesis of various related ideas than is currently available in the literature.

General $n \times m$ matrices can be considered as points in an nm-dimensional vector space. They form a group under addition, with the zero-matrix as the identity. A natural measure of distance between such matrices is the usual L_2 norm of the difference of the corresponding vectors, i.e. the square-root of the sum of the squares of the elements of the matrix given by subtracting one matrix from the other. This is known as the Frobenius norm of a matrix, and can expressed in the following equivalent formulations:

$$\begin{aligned} \|M\|_F^2 &= \|\mathrm{vec}\,(M)\|^2 \\ &= \mathrm{tr}\left(M^T M\right) = \mathrm{tr}\left(M M^T\right) \\ &= \Sigma_i s_i^2, \end{aligned}$$

where s_i are the singular values of M (see appendix A.2). The mean of several general matrices is simply the usual arithmetic mean, which can be easily shown to minimise the sum of squared distances of each matrix from the average [56].

Jacobian matrices with positive determinant[11] form a group under multiplication, with the identity matrix as the group identity. In particular, they are an example of a matrix Lie group, with an associated Lie algebra [56]. The space of matrices with positive determinant cannot be considered Euclidean, and the notions of distance and mean must hence be appropriately redefined.[12]

The special case of a 1×1 matrix with positive determinant is simply a positive scalar. For positive numbers, the fact that the most natural group is multiplicative instead of additive, suggests that in terms of distance, 0.5 and 2 are both equally far from 1 (their geometric mean), and that 0 is essentially infinitely far from any positive number. This corresponds to a distance metric

$$\begin{aligned} d_{\log}(x, y) &= |\log(y/x)| = |\log(x/y)| \\ &= |\log y - \log x| = |\log x - \log y|. \end{aligned}$$

[11] Note that the Jacobian is not positive definite, as mistakenly stated in [23] (p.131); it can have complex eigenvalues, or repeated negative-real eigenvalues while retaining a positive determinant and hence corresponding to an invertible transformation.

[12] The situation is similar to that of diffeomorphisms; as mentioned briefly in section 1.5.1, the diffeomorphism group can be seen as an infinite dimensional analogue of a Lie group.

The geometric mean is then the usual arithmetic mean in log-space,

$$\mu = \left(\prod_{i=1}^{n} p_i\right)^{1/n}$$
$$\log \mu = \frac{1}{n}\sum_{i=1}^{n} \log p_i,$$

and can be shown to minimise the sum of squared distances from itself, where (hyperbolic [57]) distances are measured according to the metric $d_{\log}$. The logarithmic distance can easily be shown to satisfy the conditions of a metric: symmetry, $d_{\log}(x, y) = d_{\log}(y, x)$; positive-definiteness, $d_{\log}(x, y) \geq 0$, $d_{\log}(x, y) = 0 \Leftrightarrow x = y$; and the triangle inequality,

$$\begin{aligned} d_{\log}(a, b) + d_{\log}(b, c) &= |\log b - \log a| + |\log c - \log b| \\ &\geq |\log b - \log a + \log c - \log b| = |\log c - \log a| = d_{\log}(a, c). \end{aligned}$$

This metric also satisfies some additional desirable properties: invariance to scaling, $d_{\log}(ax, ay) = d_{\log}(x, y)$; and invariance to inversion $d_{\log}(1/x, 1/y) = d_{\log}(x, y)$.

Distances between Jacobian matrices

Given the logarithmic distance metric for positive numbers, it might seem intuitive to attempt to use the matrix logarithm (appendix B) to define a similar distance between matrices with positive determinant like the Jacobian tensors. For example,

$$d_{\text{logm}}(X, Y) = \|\text{logm}\left(X^{-1}Y\right)\|_F = \|\text{logm}\left(Y^{-1}X\right)\|_F.$$

(Equality of the two expressions here arises from equation (B.11).)

In fact, for the special case of rotation matrices, such a distance does provide a valid metric, satisfying several useful properties. Moakher [57] defines the Riemannian distance between two rotations as $d_{rot}(R_1, R_2) = \|\text{logm}\left(R_1^T R_2\right)\|_F/\sqrt{2}$, which is simply a scaled version of d_{logm}, since $R^T = R^{-1}$ for rotations. As for $d_{\log}$, this metric is also invariant under inversion, and (now that $AX \neq XA$) invariant to both left and right multiplication by rotation matrices, i.e. it is bi-invariant: $d_{rot}(R_3R_1R_4, R_3R_2R_4) = d_{rot}(R_1, R_2)$ [57].

Unfortunately, with more general matrices, d_{logm} is not a valid metric. For the matrix logarithm, the failure of commutation in the general case (cf. equation B.13) means $\text{logm}\left(X^{-1}Y\right) \neq \text{logm}\left(YX^{-1}\right) \neq \text{logm}\,(Y) - \text{logm}\,(X)$, and nor are their Frobenius norms equal. For rotations, invariance of the Frobenius norm to matrix similarity (and hence congruence given $R^{-1} = R^T$) means that

$$\begin{aligned} \|\text{logm}\left(R_1^T R_2\right)\|_F &= \|R_1 \text{logm}\left(R_1^T R_2\right) R_1^T\|_F \\ &= \|R_1 \text{logm}\left(R_1^T R_2\right) R_1^{-1}\|_F \\ &= \|\text{logm}\left(R_1 R_1^T R_2 R_1^{-1}\right)\|_F \\ &= \|\text{logm}\left(R_2 R_1^T\right)\|_F. \end{aligned}$$

However, neither expression is equal to $\|\mathrm{logm}\,(R_2) - \mathrm{logm}\,(R_1)\|_F$.

The expression, $\|\mathrm{logm}\,(Y) - \mathrm{logm}\,(X)\|_F$, can immediately be seen to satisfy the triangle inequality thanks to the equivalence of the Frobenius norm to the standard vector norm, but it can be verified that d_{logm} does not satisfy the triangle inequality for general matrices with positive determinant (or even the stricter class of positive definite matrices). To provide a sufficient counter-example, the following MATLAB code generates three pseudo-random symmetric positive definite matrices,

```
randn('state', 0); % seed the random number generator
A = randn(3); A = A'*A;
B = randn(3); B = B'*B;
C = randn(3); C = C'*C;
```

for which

$$d_{\mathrm{logm}}(A, B) + d_{\mathrm{logm}}(B, C) > d_{\mathrm{logm}}(A, C),$$

as required, but

$$d_{\mathrm{logm}}(C, A) + d_{\mathrm{logm}}(A, B) < d_{\mathrm{logm}}(C, B),$$

implying $d_{\mathrm{logm}}(C, B)$ is not the (shortest) distance between C and B, and hence that d_{logm} is not a distance metric.

Interestingly, it can be shown that there is in fact no valid bi-invariant Riemannian metric for general matrices with positive determinant [56].[13] This implies that a bi-invariant Fréchet mean of Jacobian matrices is not a well-defined concept. We will return to this in section 4.5.

Euclidean analysis of Jacobian tensors and determinants

Log-transformation of the determinant of the Jacobian in classical univariate TBM is motivated by both statistical arguments (improved normality of the unbounded values) and Riemannian ones (conformance with the group structure). However, one can clearly still analyse the determinant without the log-transformation. Spatial smoothing, if desired, will preserve the positivity of the determinant, and standard test statistics can be applied, though they are not expected to be optimal.

It might therefore seem that one could also ignore the Riemannian argument in the case of multivariate TBM of the full Jacobian matrix, and simply analyse the tensor as a general matrix. However, somewhat surprisingly, major problems can theoretically arise with such an approach. For symmetric positive-definite matrices (considered in detail next) half the sum of two matrices is not a good Riemannian average, but it does nevertheless give a valid symmetric positive definite matrix, since $x^T(A + B)x = x^T Ax + x^T Bx > 0$ if A and B are positive definite, and $A + B$ clearly remains symmetric if both A and B are. One might similarly expect that a naïve Euclidean mean of Jacobian matrices would still be a matrix with positive determinant. However, this turns out not to be the case. A simple counter-example can be generated in MATLAB:

[13] Woods [56] goes into greater detail, but essentially, all 'compact' Lie groups have bi-invariant metrics, some non-compact ones may, but the group of matrices with positive determinant does not.

```
seed = 3; % found through (short) search of 1,2,...
randn('state', seed)
% Generate two matrices with positive determinant
J1 = randn(3); J1 = J1 / sign(det(J1));
J2 = randn(3); J2 = J2 / sign(det(J2));
det(J1 + J2) % the determinant of their sum is negative!
```

In fact, even if one is stricter still with the creation of the Jacobian matrices, and repeatedly calls `randn(3)` until achieving matrices with positive determinant and with no eigenvalues on the negative real line (see section 4.5) it is still possible (`seed=5`) for the determinant of their sum to be negative. However, this is a theoretical problem, which might not occur in practice for TBM — particularly for longitudinal studies, where the Jacobians tend to be relatively close to the identity. Studholme and Cardenas [24] have in fact published a simple Euclidean analysis of Jacobian matrices from longitudinal deformations, ignoring the complexities discussed here. For the sake of comparison, we therefore also include this option in our experimental investigations, but, importantly, we check whether negative determinants occur during the smoothing or statistical analysis. Note that although the method is theoretically unappealing, it could potentially turn out to be powerful in practice; see the related discussion of log-Euclidean analysis of strain tensors in section 4.3.6.

4.2.7 A Riemannian metric for symmetric positive definite matrices

Matrices that are symmetric positive definite (SPD) also lie on a non-Euclidean manifold. Since 2×2 symmetric matrices have only three unique elements, it is possible to visualise this manifold in three dimensional space. SPD matrices lie in the interior of a cone [58],[14] whose surface, for the 2×2 case, is defined by the equation

$$\begin{vmatrix} x & z \\ z & y \end{vmatrix} = xy - z^2 = 0,$$

together with $x > 0$ and $y > 0$. Figure 4.2 illustrates this surface.

Symmetric matrices do not form a group under multiplication, because AB is not necessarily symmetric. Real SPD matrices have unique real SPD square roots and inverses (see appendix sec:sqrtm), so the operation $A \circ B = A^{1/2} B A^{1/2}$, produces SPD results. This operation features the usual matrix identity as its identity element and the matrix inverse as its inverse element: $A \circ A^{-1} = I = A^{-1} \circ A$. However, it does not yield a group, because the axiom of associativity is not satisfied: $A \circ (B \circ C) \neq (A \circ B) \circ C$ [59].

SPD matrices nevertheless do lie on a manifold for which a Riemannian metric exists. Replacing the term $X^{-1}Y$ with the similar[15] term $X^{-1} \circ Y = X^{-1/2} Y X^{-1/2}$, converts d_{logm} to:

$$d_{\text{Aff}}(X, Y) = \|\text{logm}\left(X^{-1/2} Y X^{-1/2}\right)\|_F. \tag{4.6}$$

[14] Technically, a half-cone, and a hyperdimensional one for matrices larger than 2×2.

[15] These two matrices are 'similar' in the technical sense because they are related by $X^{-1/2} Y X^{-1/2} = C X^{-1} Y C^{-1}$ for $C = X^{1/2}$, and hence have the same eigenvalues.

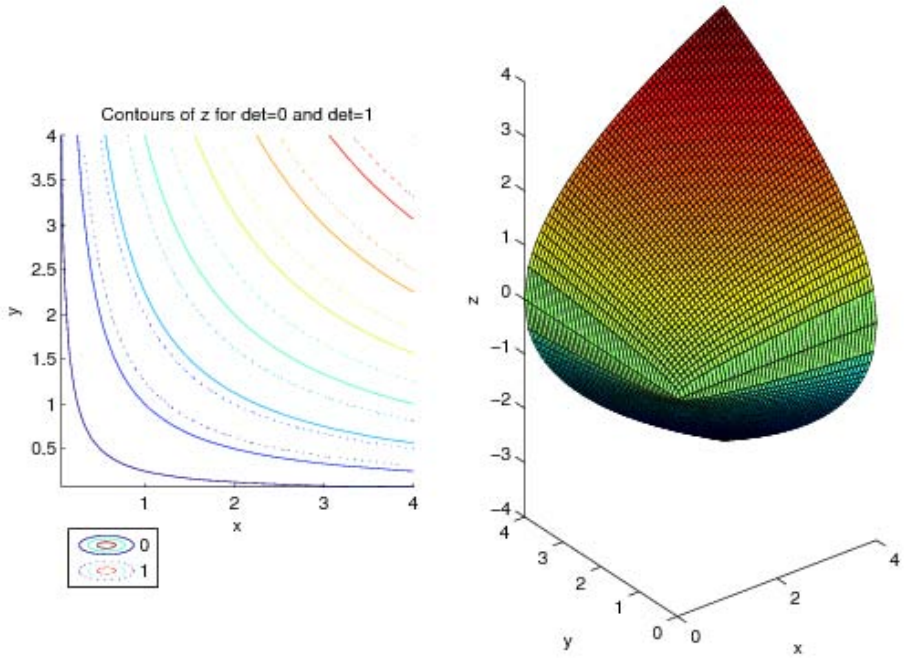


Figure 4.2: Visualisation of the cone of 2×2 SPD matrices in 3-D space. Right: the surface of the cone defined by $z = \pm\sqrt{xy}$. Note that matrices *on* this surface are not positive definite (they have determinant 0), while those inside the cone are. Left: z-contours of this surface (solid lines) and of an interior surface (for SPD matrices with unity determinant; broken lines) are plotted in the x-y plane.

It is informative to show the symmetry of the above metric, which is not obvious since $\operatorname{logm}\left(X^{-1/2}\,Y\,X^{-1/2}\right) \neq \operatorname{logm}\left(Y^{-1/2}\,X\,Y^{-1/2}\right)$ in general. Noting that the logarithm preserves the symmetry of $X^{-1/2}\,Y\,X^{-1/2}$ and using the matrix similarity (B.8)

$$\operatorname{logm}\left(X^{-1/2}\,Y\,X^{-1/2}\right) = X^{1/2}\operatorname{logm}\left(X^{-1}Y\right)X^{-1/2},$$

followed by the trace's circularity-property to reduce $\operatorname{tr}\left(X^{1/2}\cdots X^{-1/2}\right) = \operatorname{tr}\left(\cdots\right)$, we have:

$$\begin{aligned}
\|\operatorname{logm}\left(X^{-1/2}\,Y\,X^{-1/2}\right)\|_F^2 &= \operatorname{tr}\left(\operatorname{logm}\left(X^{-1/2}\,Y\,X^{-1/2}\right)^T \operatorname{logm}\left(X^{-1/2}\,Y\,X^{-1/2}\right)\right) \\
&= \operatorname{tr}\left(\operatorname{logm}\left(X^{-1/2}\,Y\,X^{-1/2}\right)\operatorname{logm}\left(X^{-1/2}\,Y\,X^{-1/2}\right)\right) \\
&= \operatorname{tr}\left(X^{1/2}\operatorname{logm}\left(X^{-1}Y\right)X^{-1/2}\,X^{1/2}\operatorname{logm}\left(X^{-1}Y\right)X^{-1/2}\right) \\
&= \operatorname{tr}\left(\operatorname{logm}\left(X^{-1}Y\right)\operatorname{logm}\left(X^{-1}Y\right)\right).
\end{aligned}$$

At this point, it is important to note that

$$\operatorname{tr}\left(\operatorname{logm}\left(X^{-1}Y\right)\operatorname{logm}\left(X^{-1}Y\right)\right) \neq \operatorname{tr}\left(\operatorname{logm}\left(X^{-1}Y\right)^T\operatorname{logm}\left(X^{-1}Y\right)\right) = \|\operatorname{logm}\left(X^{-1}Y\right)\|_F^2$$

due to the lack of symmetry of $\operatorname{logm}\left(X^{-1}Y\right)$. Continuing the derivation, following the

reverse of the above sequence but with Y in place of X gives:

$$\begin{aligned}\|\operatorname{logm}\left(X^{-1/2}\,Y\,X^{-1/2}\right)\|_F^2 &= \operatorname{tr}\left(\operatorname{logm}\left(X^{-1}Y\right)\operatorname{logm}\left(X^{-1}Y\right)\right)\\ &= \operatorname{tr}\left(Y^{1/2}\operatorname{logm}\left(X^{-1}Y\right)Y^{-1/2}\,Y^{1/2}\operatorname{logm}\left(X^{-1}Y\right)Y^{-1/2}\right)\\ &= \operatorname{tr}\left(\operatorname{logm}\left(Y^{1/2}\,X^{-1}Y^{1/2}\right)\operatorname{logm}\left(Y^{1/2}\,X^{-1}Y^{1/2}\right)\right)\\ &= \operatorname{tr}\left(\operatorname{logm}\left(Y^{1/2}\,X^{-1}Y^{1/2}\right)^{T}\operatorname{logm}\left(Y^{1/2}\,X^{-1}Y^{1/2}\right)\right)\\ &= \|\operatorname{logm}\left(Y^{1/2}\,X^{-1}Y^{1/2}\right)\|_F^2.\end{aligned}$$

Finally, using the invariance of the norm to sign, and (B.11) shows the symmetric form:

$$\begin{aligned}\|\operatorname{logm}\left(X^{-1/2}\,Y\,X^{-1/2}\right)\|_F^2 &= \|-\operatorname{logm}\left(Y^{1/2}\,X^{-1}Y^{1/2}\right)\|_F^2\\ &= \|\operatorname{logm}\left(\left(Y^{1/2}\,X^{-1}Y^{1/2}\right)^{-1}\right)\|_F^2\\ &= \|\operatorname{logm}\left(Y^{-1/2}\,XY^{-1/2}\right)\|_F^2.\end{aligned}$$

The metric d_{Aff} has been derived in alternative ways by Batchelor et al. [60] and Pennec et al. [58]; the first authors start with an expression for the length of an arc between an identity and an infinitesimally close tensor, deriving the geodesic between the identity and a diagonal matrix, and then using congruence (see below) to build up to the distance between two general SPD tensors. The latter paper also uses congruence to simplify $d(X,Y)$ to $d(I, X^{-1/2}\,YX^{-1/2})$, and then uses invariance arguments to show that the distance should be a function of the squared logs of the eigenvalues of $X^{-1/2}\,YX^{-1/2}$; the Frobenius norm of the matrix logarithm of $X^{-1/2}\,YX^{-1/2}$ satisfies this requirement because, for an SPD matrix A, $\|A\|_F^2 = \operatorname{tr}\left(A^2\right)$ and equations (B.2) and (B.6) respectively imply that this equals the sum of squared eigenvalues of A, which for $A = \operatorname{logm}(B)$ are the logs of the eigenvalues of B.

Under a general linear change of coordinates $x \to Ax$, the Jacobian matrix changes as $J \to AJA^{-1}$ and therefore the SPD strain tensors change as $JJ^T \to AJJ^TA^T$ or $J^TJ \to BJ^TJB^T$ where $B = A^{-T}$, i.e. a congruence relation, as for the diffusion tensor [58, 60]. The metric d_{Aff} is invariant to congruence transformations by definition, i.e. $d_{\text{Aff}}(AXA^T, AYA^T) = d_{\text{Aff}}(X,Y)$ for any invertible A. To see this, note that

$$d_{\text{Aff}}(AXA^T, AYA^T) = \|\operatorname{logm}\left((AXA^T)^{-1/2}(AYA^T)(AXA^T)^{-1/2}\right)\|_F$$

is the Frobenius norm of the logarithm of an SPD matrix, and therefore depends on its eigenvalues, as shown above. Now, using similarity, the following matrices have the same

eigenvalues:

$$\begin{aligned} D &= (AXA^T)^{-1/2}(AYA^T)(AXA^T)^{-1/2} \\ &\sim (AXA^T)^{-1/2}D(AXA^T)^{1/2} \\ &= (AXA^T)^{-1}AYA^T = A^{-T}X^{-1}YA^T \\ &\sim X^{-1}Y \sim X^{-1/2}\,YX^{-1/2}. \end{aligned}$$

The last of the above matrices is SPD, and hence its eigenvalues provide its Frobenius norm, and in turn the original distance $d_{\mathrm{Aff}}(X,Y)$.

This invariance under a linear transformation also ensures affine invariance [58], since translations affect neither the diffusion nor Jacobian tensors. This is the origin of the nomenclature d_{Aff}, which is adopted from [59].

The metric is also invariant under inversion

$$\begin{aligned} d_{\mathrm{Aff}}(X^{-1},Y^{-1}) &= \|\mathrm{logm}\left(X^{1/2}\,Y^{-1}X^{1/2}\right)\|_F \\ &= \|\mathrm{logm}\left((X^{1/2}\,Y^{-1}X^{1/2})^{-1}\right)\|_F \\ &= \|\mathrm{logm}\left(X^{-1/2}\,YX^{-1/2}\right)\|_F = d_{\mathrm{Aff}}(X,Y). \end{aligned}$$

The combination of affine invariance and invariance to inversion can also be interpreted as a form of bi-invariance [56], though this should not be confused with the suggestion that $d_{\mathrm{Aff}}(PXQ,PYQ) = d_{\mathrm{Aff}}(X,Y)$ which is only true for $Q = P^T$.

This affine invariant metric can be used to define a Fréchet mean of SPD tensors T_i which satisfies $\sum_i \mathrm{logm}\left(T_i^{-1/2}\bar{T}T_i^{-1/2}\right) = 0$.[16] Statistical tests can be carried out within the tangent plane to the Riemannian manifold at the location of this mean[58]. However, the computations are expensive, especially when they must be performed at every voxel for a set of high-resolution structural MR images. For example to find an explicit estimate of the implicitly defined mean, it is necessary to perform a geodesic gradient descent [58]:

$$\begin{aligned} \bar{T}_{t+1} &= \bar{T}_t^{1/2}\mathrm{expm}\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{logm}\left(\bar{T}_t^{-1/2}T_i\bar{T}_t^{-1/2}\right)\right)\bar{T}_t^{1/2} \\ &= \bar{T}_t\,\mathrm{expm}\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{logm}\left(\bar{T}_t^{-1}T_i\right)\right) \\ &= \mathrm{expm}\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{logm}\left(T_i\bar{T}_t^{-1}\right)\right)\bar{T}_t. \end{aligned}$$

Where the equality of the above expressions follows from:

$$\begin{aligned} \bar{T}_t^{1/2}\mathrm{expm}\,(Z)\,\bar{T}_t^{1/2} &= \bar{T}_t^{1/2}\mathrm{expm}\,(Z)\,\bar{T}_t^{-1/2}\bar{T}_t = \mathrm{expm}\left(\bar{T}_t^{1/2}Z\bar{T}_t^{-1/2}\right)\bar{T}_t \\ &= \bar{T}_t\bar{T}_t^{-1/2}\mathrm{expm}\,(Z)\,\bar{T}_t^{1/2} = \bar{T}_t\mathrm{expm}\left(\bar{T}_t^{-1/2}Z\bar{T}_t^{1/2}\right), \end{aligned}$$

[16] This expression is equivalent to $\sum_i \mathrm{logm}\left(\bar{T}^{-1/2}T_i\bar{T}^{-1/2}\right) = 0$ and also to $\sum_i \mathrm{logm}\left(T_i^{-1}\bar{T}\right) = 0$ which is the form given in [60] (see the discussion of the equivalence of the gradient descent formulae).

bringing the matrix similarities inside the exponential, and then in turn, inside each logarithm within the sum (i.e. using $C\left(\sum_{i=1}^{N}\operatorname{logm}(D)\right)C^{-1}=\sum_{i=1}^{N}\operatorname{logm}\left(CDC^{-1}\right)$).

The Log-Euclidean Riemannian metric

A more computationally efficient metric is available if one is willing to sacrifice affine invariance. For scalars, d_{Aff} reduces to $d_{\log}$, just as d_{logm} does. Now, it was observed above that while $d_{\log}(x,y)=|\log(y/x)|=|\log y-\log x|$, for matrices

$$\begin{aligned}d_{\text{logm}}(X,Y)&=\|\operatorname{logm}\left(X^{-1}Y\right)\|_F\\&\neq\|\operatorname{logm}(Y)-\operatorname{logm}(X)\|_F=d_{\text{LE}}(X,Y).\end{aligned}$$

Now, unlike d_{logm}, d_{LE} is a generalisation of $d_{\log}$ which does satisfy the triangle inequality for SPD tensors.[17] In fact, d_{LE}, which was first proposed and extensively developed by Arsigny et al. [59, 61], satisfies all the necessary properties of a metric, and is additionally invariant under inversion. Considering a congruence, the metric is invariant to scalar multiplication thanks to (B.17), and for the special case of an orthogonal matrix $P^T=P^{-1}$ allows us to use (B.8) to show

$$\begin{aligned}d^2_{\text{LE}}(PXP^T,PYP^T)&=\|\operatorname{logm}\left(PYP^T\right)-\operatorname{logm}\left(PXP^T\right)\|_F^2\\&=\|P\left\{\operatorname{logm}(Y)-\operatorname{logm}(X)\right\}P^T\|_F^2=\|PQP^T\|_F^2\\&=\operatorname{tr}\left((PQP^T)^TPQP^T\right)=\operatorname{tr}\left(PQ^TP^TPQP^T\right)\\&=\operatorname{tr}\left(PQ^TQP^T\right)=\operatorname{tr}\left(P^TPQ^TQ\right)=\operatorname{tr}\left(Q^TQ\right)\\&=\|Q\|_F^2=\|\operatorname{logm}(Y)-\operatorname{logm}(X)\|_F^2=d^2_{\text{LE}}(X,Y).\end{aligned}$$

So the metric is invariant to congruence by a geometric similarity transformation.[18] It is simple to verify in practice that it is not invariant to more general affine transformations such as anisotropic scaling or skewing.

The key benefit of this metric is that it provides a Euclidean vector space structure on tensors, hence the name 'log-Euclidean' which d_{LE} denotes. This aspect is described in much greater detail in [59], but we shall briefly describe some important properties here. By defining a vectorisation operator that extracts the n diagonal elements and the $n(n-1)/2$ unique off-diagonal elements, dividing the off-diagonal terms by $\sqrt{2}$, the Frobenius norm maps to the standard L_2 norm of the vectors:

$$\|\operatorname{logm}(Y)-\operatorname{logm}(X)\|_F=\|\operatorname{vech}_{LE}(\operatorname{logm}(Y))-\operatorname{vech}_{LE}(\operatorname{logm}(X))\|.$$

Similarly, instead of the implicit formula for the Fréchet mean based on d_{Aff}, the log-

[17] It is not clear from the literature whether it is a suitable metric for more general matrices, e.g. those with real logarithms. See section 4.5 for further discussion of this.

[18] We use the term geometric similarity to distinguish this similarity transformation (rotation and scaling) from the matrix similarity transformation used earlier.

Euclidean metric leads to a very natural explicit formula

$$\operatorname{logm}\left(\bar{T}\right) = \frac{1}{N}\sum_{i=1}^{N}\operatorname{logm}\left(T_i\right),$$

i.e. the same form as the log of the geometric mean of scalars given near the start of this section. In fact, the log-Euclidean approach automatically makes all the standard statistical methods available to SPD tensors, without any additional complication. It has been used for analysis of diffusion tensors by Whitcher et al. [16], who also compared it to straight-forward Euclidean analysis. Log-Euclidean analysis of a deformation tensor (actually the right stretch tensor U from section 4.2.5) has been explored in the work of Lepore et al. [23] on cross-sectional TBM.

The main disadvantage of the log-Euclidean metric is that the loss of affine-invariance means that the choice of template subject affects the analysis, in contrast to the framework expounded in [56]. However, for our application of longitudinal TBM, the tensors of interest relate to the chronological warps, and the choice of template for cross-sectional normalisation is of considerably less importance. For this reason, the computationally simpler log-Euclidean analysis of strain tensors is particularly appealing, and is the main method focussed on here.

We have shown that analysis of the matrix logarithms of a symmetric positive definite matrix derived from the Jacobian can be derived from two very different view-points. Solid-mechanics led to the idea of analysing the unique elements of the Hencky tensor, $\operatorname{vech}\left(\operatorname{logm}\left((J^TJ)^{1/2}\right)\right)$, as proposed by Ashburner [2]. The desire for a Riemannian metric led to $\operatorname{vech}_{LE}\left(\operatorname{logm}\left((J^TJ)\right)\right)$, or $\operatorname{vech}_{LE}\left(\operatorname{logm}\left((J^TJ)^{1/2}\right)\right)$ as analysed by Lepore et al. [23]. The matrix square-root within the matrix logarithm can be taken out as an irrelevant factor of two, leaving the only difference between these approaches being the $\sqrt{2}$ normalisation of the unique off-diagonal elements, motivated from the desire to equate the Frobenius and vector norms. In the experimental section, we investigate the practical relevance of this distinction.

4.2.8 Further Jacobian-based measures

It may also be useful to consider other measures derived from the Jacobian or strain tensors. There are several possible motivations for this: (a) although Jacobian-derived measures can only preserve or decrease statistical information, non-linearly derived measures may better exploit their information in practical significance tests; (b) reducing the dimensionality from the nine Jacobian elements (with their 45 covariances) may improve inference from limited numbers of subjects; (c) some aspects may be of less fundamental interest, for example rotational information in the Jacobian, as discussed in section 4.2.5; (d) measures that focus on particular aspects of interest, e.g. magnitudes of principal axes of strains, or orientation of the major axis, lead to easier and more precise interpretation of significant findings.

One natural idea is to consider components of an eigendecomposition. However, the

Jacobian matrices can (and often do in practice) have complex eigenvalues, which complicates both testing and interpretation. The singular values of the Jacobian matrix are guaranteed to be real, and also positive for a Jacobian with positive determinant (since they are equal to the eigenvalues of the SPD matrix $J^T J$). Their positivity leads naturally to consider their logarithms, at which point we observe that the Hencky tensor has arisen again, since the logs of the singular values of the Jacobian are the logs of the square roots of the eigenvalues of $J^T J$, which in turn are equal to the eigenvalues of $\operatorname{logm}\left(\left(J^T J\right)^{1/2}\right)$. The eigenvalues of the Hencky tensor contain all of its information about the magnitudes of the principal strains, without the information about the principal axes. Lepore et al. [23] compared analyses based on $\operatorname{logm}(U)$, the eigenvalues of U (without the log), the maximum (log) eigenvalue,[19] the log determinant (equal to $\log|J|$), and the log trace,[20] in a two-dimensional example over a slice through segmented corpus callosa. The same experiment also included two orientational measures, discussed further below.

4.2.9 Measures of vorticity, anisotropy and orientation

The focus thus far (both above, and in the literature) has been on deformation or strain. The main reason for this is probably the ease of interpretation of displacement and of volume change. However, it is possible that measurements related more to the directional or rotational aspects of the deformation might contain useful information. Patterns of significant group-difference in such measures would be more challenging to interpret, but may nevertheless be of interest. We therefore consider several less common measures, loosely grouped together as orientational. We choose to group vorticity and anisotropy with other orientational measures, even though they are not dependent on direction as such, simply because they seem more closely related to the orientational measures than to the more standard deformation or strain based measures described earlier.

Historically, the first such measure seems to have been proposed by Chung et al. [50], though they did not present results for it. Considering $(K+K^T)/2$, the infinitesimal strain tensor (section 4.2.5), they note that the complementary skew-symmetric term from the additive decomposition $K = (K + K^T)/2 + (K - K^T)/2$ encodes rotational information. This term has already been mentioned in section 4.2.5 as the infinitesimal rotation tensor [47]. Ignoring their signs, three unique elements of $K - K^T = J - J^T$ can be identified as

$$\begin{aligned} &\partial u_z/\partial y - \partial u_y/\partial z, \\ &\partial u_x/\partial z - \partial u_z/\partial x, \\ &\partial u_y/\partial x - \partial u_x/\partial y, \end{aligned}$$

which are the elements of the curl of the displacement vector field, hence forming a natural complement to the displacement's gradient (Jacobian matrix) and divergence (the dilata-

[19] This inconsistency seems surprising, and might be a mistake in the paper, but the authors explicitly refer to the log of the deformation tensor and to the eigenvalues of the deformation tensor.

[20] Note that $\operatorname{tr}(U) = \operatorname{tr}\left((J^T J)^{1/2}\right)$ is related to a root-mean-square of the eigenvalues of the Jacobian matrix, instead of the simpler mean that relates to $\operatorname{tr}(J)$.

tion, closely related to the Jacobian determinant). Note that both curl and the Jacobian itself are linear (differential) operators, and Gaussian smoothing is also a linear operation, so the curl of the Jacobian of the smoothed displacement field is the same as the curl of the smoothed Jacobian and the smoothed curl.

Strain anisotropy

Moving from the rotational, vorticity aspect of the displacement, to a more directional aspect of the three-dimensional deformation, we may wish to investigate the anisotropy of the strain; i.e. to what extent is the ellipsoid of the strain tensor preferentially elongated, or, informally, how 'oriented' is it? The fractional anisotropy (FA) is a measure related to the second moment (or more precisely, the sample standard deviation) of the eigenvalues [62]. For comparison, the trace of an $m \times m$ matrix is simply m times the arithmetic mean or first moment of its eigenvalues, which is known as the mean-diffusivity in diffusion tensor imaging [62]. Higher moments of the distribution of eigenvalues, such as skewness, may also be of interest [63]. In 3D, FA can be expressed in terms of the eigenvalues $\{\lambda_i\}_{i=1}^3$ and their mean $\bar{\lambda} = \frac{1}{3}\mathrm{tr}\,(T)$, as

$$\begin{aligned}
\mathrm{FA} &= \frac{\sqrt{3}}{\sqrt{2}} \frac{\sqrt{\sum_{i=1}^3 (\lambda_i - \bar{\lambda})^2}}{\sqrt{\sum_{i=1}^3 \lambda_i^2}} \\
&= \frac{\sqrt{3}}{\sqrt{2}} \frac{\|T - I\bar{\lambda}\|_F}{\|T\|_F} \\
&= \frac{\sqrt{3}}{\sqrt{2}} \frac{d(T, I\,\mathrm{tr}\,(T)\,/3)}{d(T, 0)}
\end{aligned}$$

The final expression directly shows the relation of FA to how far the strain ellipsoid deviates from being spherical, or how far the strain tensor is from the closest isotropic tensor, using distance measured with a Euclidean metric.[21] Although not novel, the maths demonstrating that $I\,\mathrm{tr}\,(T)\,/m$ is the closest scaled identity matrix to an $m \times m$ tensor T is not included in [60], so for completeness it is given here. The (squared) distance from T to a general scaled identity matrix is

$$\begin{aligned}
\|T - tI\|_F^2 &= \mathrm{tr}\left((T - tI)^2\right) \qquad \text{(noting } T - tI \text{ is symmetric)} \\
&= \mathrm{tr}\left(T^2 + t^2 I - 2tT\right) \\
&= \mathrm{tr}\left(T^2\right) + mt^2 - 2t\mathrm{tr}\,(T)\,,
\end{aligned}$$

and the minimum requires stationarity

$$\begin{aligned}
\frac{d\|T - tI\|_F^2}{dt} = 0 \Rightarrow 2mt - 2\mathrm{tr}\,(T) &= 0 \\
t &= \mathrm{tr}\,(T)\,/m.
\end{aligned}$$

[21] At this point, recall the similar definition of the deviatoric infinitesimal strain tensor in section 4.2.5.

Following the same logic as section 4.2.6, Batchelor et al. [60] proposed to use an affine-invariant Riemannian distance metric in place of Euclidean distance, giving the geodesic anisotropy (GA). Note that the normalising denominator $d(T,0)$ and constants of proportionality in the definition of FA ensure that $0 \leq \mathrm{FA} \leq 1$; however, $d_{\mathrm{Aff}}(T,0) = \infty$, so Batchelor et al. instead chose to define an unnormalised measure satisfying $0 \leq \mathrm{GA} < \infty$ and to normalise it as $0 \leq \tanh(\mathrm{GA}) < 1$. We have:

$$\mathrm{GA} = \min_t d_{\mathrm{Aff}}(T, tI) = \min_t d^2_{\mathrm{Aff}}(T, tI),$$

where the squared distance can be expanded as

$$\begin{aligned} d^2_{\mathrm{Aff}}(T,tI) &= \|\mathrm{logm}\left(t^{-1/2}Tt^{-1/2}\right)\|_F^2 = \|\mathrm{logm}\left(T/t\right)\|_F^2 \\ &= \|\mathrm{logm}\left(T(tI)^{-1}\right)\|_F^2 = \|\mathrm{logm}\left(T\right) - \mathrm{logm}\left(tI\right)\|_F^2 = d^2_{\mathrm{LE}}(T,tI) \\ &= \mathrm{tr}\left(\left(\mathrm{logm}\left(T\right) - \mathrm{logm}\left(tI\right)\right)^2\right) \\ &= \mathrm{tr}\left(\mathrm{logm}^2\left(T\right) + \mathrm{logm}^2\left(tI\right) - 2\mathrm{logm}\left(T\right)\mathrm{logm}\left(tI\right)\right) \\ &= \mathrm{tr}\left(\mathrm{logm}^2\left(T\right) + I\log^2(t) - 2\log(t)\mathrm{logm}\left(T\right)\right) \\ &= \mathrm{tr}\left(\mathrm{logm}^2\left(T\right)\right) + m\log^2(t) - 2\log(t)\mathrm{tr}\left(\mathrm{logm}\left(T\right)\right). \end{aligned} \tag{4.7}$$

The minimum requires

$$\begin{aligned} \frac{d^2_{\mathrm{Aff}}(T,tI)}{d\log t} = 0 \Rightarrow \log t &= \tfrac{1}{m}\mathrm{tr}\left(\mathrm{logm}\left(T\right)\right) \\ \log t = \tfrac{1}{m}\log|T| &= \log(|T|^{1/m}) \Rightarrow t = |T|^{1/m} \\ \mathrm{GA} &= d_{\mathrm{Aff}}(T, I|T|^{1/m}) = d_{\mathrm{LE}}(T, I|T|^{1/m}). \end{aligned} \tag{4.8}$$

Note that (4.7) proves that the affine-invariant and log-Euclidean metrics lead to equivalent measures of geodesic anisotropy, using the fact that the isotropic tensor commutes with the original one, as observed by Lepore et al. [23]. We note, novelly to the best of our knowledge, that the equivalence of GA under affine-invariant and log-Euclidean metrics also shows that the GA shares the FA's simple interpretation in terms of the sample standard deviation of the eigenvalues [62], since both FA and GA relate to $\|T - tI\|_F$ where T is either the SPD tensor itself (FA) or its matrix logarithm (GA), and $t = \mathrm{tr}\left(T\right)/3 = \bar{\lambda}(T)$. We derive the relationship for completeness:

$$\begin{aligned} \|T - tI\|_F^2 &= \mathrm{tr}\left((T-tI)^2\right) \\ &= \sum_i \lambda_i\left((T-tI)^2\right) \\ &= \sum_i \left(\lambda_i\left(T - tI\right)\right)^2 \qquad \text{Using (B.2)} \\ &= \sum_i \left(\lambda_i(T) - t\right)^2, \end{aligned}$$

where the last line uses the fact that if λ is an eigenvalue of T then $Tv = \lambda v$ gives

$(T - tI)v = (\lambda - t)v$ and so $\lambda - t$ is an eigenvalue of $T - tI$. Finally, we have

$$GA = \sqrt{\sum_i \left(\lambda_i(T) - \bar{\lambda}(T)\right)^2}, \tag{4.9}$$

which is simply $\sqrt{2}$ times the unbiased estimate of the standard deviation of the eigenvalues, or $\sqrt{3}$ times the biased estimate, as given in equation (12) of [62].

The nearest isotropic tensor under the Euclidean norm preserves the trace of the tensor, $\mathrm{tr}\left(I\,\mathrm{tr}\left(T\right)/m\right) = \mathrm{tr}\left(T\right)$; the nearest under either of the Riemannian norms preserves its determinant, $\det\left(I\left|T\right|^{1/m}\right) = \left|T\right|$.[22] This is a special case of the properties for interpolation of traces or determinants under interpolation of tensors [60].

Directionality of strain

If there is notable anisotropy in some brain regions, then there may also be regional patterns in the major axis of the strain ellipsoid, either involving its magnitude and direction jointly, or purely in terms of either one of these aspects. The magnitude is given by the largest of the eigenvalues; the orientation, known as the principal direction,[23] is given by the corresponding eigenvector. It is interesting to consider how this might compare to the direction of the displacement field itself at the same voxel.[24] One key distinction is that unlike the displacement vectors, the principal strain vectors (like the principal diffusion directions) do not have a well-defined sense, in that the three ellipsoidal axes can not be assigned positive or negative directions, meaning that $\hat{v}$ and $-\hat{v}$ are equivalent.[25] This can be seen from the nature of the principal direction as an eigenvector: if $Tv = \lambda v$ then trivially $-Tv = -\lambda v$ gives $T(-v) = \lambda(-v)$. Another potentially important difference is that the principal eigenvector may be poorly defined, and hence unstable in the presence of noise, if the first two or three eigenvalues are of similar magnitude (i.e. the tensor ellipsoid is oblate or spherical, rather than prolate).

The direction of a vector (or axis) is a classic example of a measurement where the manifold structure is particularly important. Lepore et al. [23] consider the principal strain direction in two dimensions (where it has a single degree of freedom) as a single angle relative to the horizontal, on which they perform standard univariate statistics. Even with the simplified manifold structure in 2D, this is still questionable. For example, the equivalence of 0 and 180° means that an angle of 10° is as close to 170° as it is to 30° — a fact not accounted for in the student's t-test used by Lepore et al. [23] which may partially explain the poor results they observe (their figures 2 and 3 show very little significance for angular differences). In 3D the direction of a vector has two degrees of freedom, and can be identified with points on the unit sphere. However, it would be highly inappropriate to perform standard statistics on the usual polar angles, not only because

[22] We denote the first determinant with $\det\left(T\right)$ here to avoid the confusing expression $\left|I\left|T\right|^{1/m}\right|$.

[23] We use the plural terms principal strains or principal axes to refer to the set of eigenvectors, and the singular principal strain or principal direction to refer to the largest of these.

[24] Such a comparison can be found later in figure 4.39.

[25] Strictly, the principal direction would be known as the principal *axis* in the field of directional statistics, because of this irrelevance of the unit vector's sense.

of the above-mentioned equivalence of opposite points, but more importantly because the actual angle between two points (which is the Riemannian distance between them) is not represented by the azimuthal angle at non-zero latitudes. For example, at latitudes near $90°$, large changes in the azimuth represent very small actual distances between the unit vectors.

Addressing the related problem in diffusion tensor imaging, Schwartzman et al. [64] used the bipolar Watson distribution to analyse the principal direction in a way that accounts for the curved manifold it inhabits. The bipolar Watson statistic is invariant to the arbitrary sense of the vectors, as desired. The hypothesis testing in [64] employed a parametric approximation, followed by false discovery rate correction. In our experimental work, we use the same permutation testing framework developed for the Cramér statistic to derive FWE-corrected p-values for a Watson-based statistic. Further details are given in section 4.3.3.

Note that the question of which particular strain tensor to base the principal direction upon does not arise, since the operations of matrix logarithm and matrix powers (including square-root) only modify the eigenvalues and not the eigenvectors. Any monotonic function of the eigenvalues will preserve their ordering and hence choice of principal eigenvector. However, the question of what quantity to smooth becomes crucial. Smoothing a field of unit vectors will not preserve their unit norm. Re-normalising smoothed unit vectors is unappealing, since a relatively rough field could include many points where smoothing leads to near cancellation of the components, requiring large scaling to re-normalise, and consequently producing noisy 'smoothed' results. Instead, it seems preferable to smooth a better behaved tensor, and then to compute the principal direction from this. The Hencky tensor seems a natural choice, given the theoretical underpinnings of log-Euclidean smoothing.

Since the principal direction unit vectors require special statistics, and also seem to be surprisingly noisy, it is interesting to consider whether a non-unit magnitude vector could have advantages. We propose one such quantity: a vector of the absolute values of the principal eigenvector components, scaled to have the magnitude of the corresponding principal eigenvalue. The result will no longer be suitable for the bipolar Watson test, but should be approximately valid for conventional statistics. Though strictly, the positivity of the components means that the vector again lies in a manifold whose structure should ideally be accounted for in the analysis. This measure is a compromise between a more purely directional quantity and the strain based measures that form the main part of this chapter; it might be hoped that such a combined measure would be more sensitive to complex patterns of group difference than either type of measure alone. In addition, because this non-unit vector measure removes the need for re-normalisation, conventional smoothing of the derived result becomes possible, as an alternative to deriving the result from the smoothed tensors.

4.2.10 Transformation of deformation fields and their derivatives

With the goal of quantifying the chronological change in multiple subjects, there are two main possibilities for processing the data: (i) processing of all images can be performed separately, with no account taken of whether they come from different subjects or different time-points within a particular subject; only the statistical analysis accounts for the longitudinal nature of the data, for example simply by analysing differences within each subject with respect to their baseline. (ii) Alternatively, longitudinal effects can be directly measured in the image processing steps, with the statistical analysis comparing these direct measures of change [65]. The second approach should have advantages when the longitudinal processing is likely to be more accurate than multiple separate applications of standard image processing. This was the motivation for the boundary shift integral approach [66] and is also likely to be the case with non-rigid registration for longitudinal VBM, DBM or TBM. The two distinct options, including multiple variations on the latter in the case of longitudinal VBM are investigated in chapter 3; here, the details of the second approach in relation to longitudinal DBM and TBM are explored.

Deformation- or tensor-based voxel-wise analysis of longitudinally processed data from multiple subjects requires the chronological transformations to be combined with inter-subject (spatial normalisation) transformations. An equivalent way of expressing this is to say that the we seek the longitudinal transformation in reference space $T_r(r)$ that corresponds to the longitudinal transformation in source space $T_s(s)$, where reference and source space are related via $s = T_{sr}(r)$. More precisely, T_r is conjugate to T_s, facilitated by T_{sr} [67]. Figure 4.3 makes it clear that for a particular point r_0 we can reach $r_1 = T_r(r_0)$ directly (using transformation d from the figure), or by composing the three transformations ($d = c \circ b \circ a$ from the figure). In detail, the four transformations a, b, c and d are respectively:

$$\begin{aligned} s_0 &= T_{sr}(r_0) \\ s_1 &= T_s(s_0) \\ r_1 &= T_{rs}(s_1) = T_{sr}^{-1}(s_1), \\ r_1 &= T_{sr}^{-1}(T_s(T_{sr}(r_0))). \end{aligned} \tag{4.10}$$

For the special case of a longitudinal displacement field $u_s(s)$, $s_1 = T_s(s_0) = s_0 + u_s(s_0)$, we have

$$\begin{aligned} u_r(r_0) = r_1 - r_0 &= T_{sr}^{-1}(T_s(T_{sr}(r_0))) - r_0 \\ &= T_{sr}^{-1}(T_{sr}(r_0) + u_s(T_{sr}(r_0))) - r_0, \end{aligned}$$

which is equivalent to the result given by Rao et al. [67]. As a visual example, in figure 4.3 it can be seen that rotation and compression of source onto reference also rotates and compresses the longitudinal change.

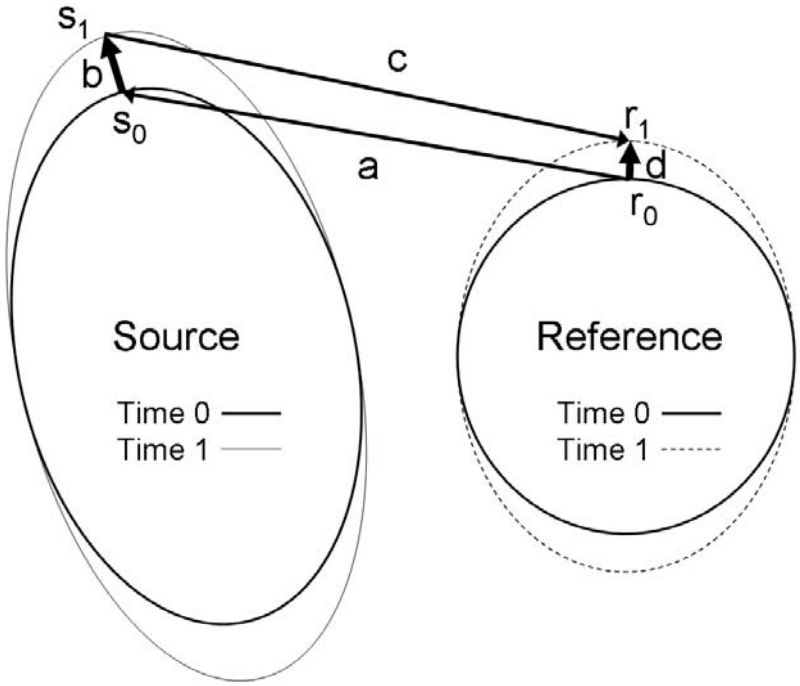


Figure 4.3: Conjugate spatial transformations. Given a mapping, a, from reference image to source image, and a transformation in source space, b, the inverse of the reference-source mapping, c, is required to derive the transformation in reference space, d, which is given by the composition $c \circ b \circ a$. Equivalently, $r_1 = d(r_0) = c(b(a(r_o)))$. Note that the time-1 reference is shown dotted since it is unknown.

Transformation of Jacobian tensor fields

Regarding the Jacobian matrices, the chain rule first gives

$$\begin{aligned} \left.\frac{\partial T_r}{\partial r}\right|_{r_0} &= \frac{\partial}{\partial r} T_{rs}(T_s(T_{sr}(r_0))) \\ &= \left.\frac{\partial T_{rs}}{\partial s}\right|_{s_1} \left.\frac{\partial T_s}{\partial s}\right|_{s_0} \left.\frac{\partial T_{sr}}{\partial r}\right|_{r_0}; \end{aligned} \tag{4.11}$$

now, using the following trick with a second application of the chain rule allows us to evaluate the Jacobian of $T_{rs} = T_{sr}^{-1}$ in terms of the Jacobian of T_{sr}:

$$\begin{aligned} s_1 &= T_{sr}(T_{rs}(s_1)) \\ \left.\frac{\partial s}{\partial s}\right|_{s_1} &= I = \left.\frac{\partial T_{sr}}{\partial r}\right|_{r_1 = T_{rs}(s_1)} \left.\frac{\partial T_{rs}}{\partial s}\right|_{s_1} \\ \left.\frac{\partial T_{rs}}{\partial s}\right|_{s_1} &= \left[\left.\frac{\partial T_{sr}}{\partial r}\right|_{r_1 = T_{rs}(s_1)}\right]^{-1} \\ J_{rs}(s_1) &= J_{sr}^{-1}(r_1), \end{aligned} \tag{4.12}$$

which mirrors the result in the appendix of [67], and makes intuitive sense in terms of the volume ratios given by the Jacobian determinants: the voxel at s_1 changes volume by the

reciprocal of the volume change of the related voxel at r_1, during the transformations T_{rs} and T_{sr} that relate them.

Finally, this allows us to write equation 4.11 as

$$J_r(r_0) = J_{sr}^{-1}(r_1)J_s(s_0)J_{sr}(r_0). \tag{4.13}$$

One of the few papers which attempts to analyse full Jacobian tensors from longitudinal registration after inter-subject spatial normalisation is that of Studholme and Cardenas [24]. However, they appear to have crucially misunderstood the work of Rao et al. [67], citing it, but dismissing it as relating only to deformation fields, and not the Jacobian matrix with which they are concerned. In fact, Rao et al. also derive results relating the Jacobian tensors of conjugate transformations, as we reiterated above. In [24], the finite strain (FS) reorientation strategy of [63] is employed. This is discussed further in section 4.2.10, but for now, we recall from [63] that FS reorientation neglects the non-rigid components of the transformation (including affine components of scaling and shearing). Figure 4.4 illustrates the behaviour of the FS method on deformation fields, which we argue is undesirable for macroscopic longitudinal deformation.

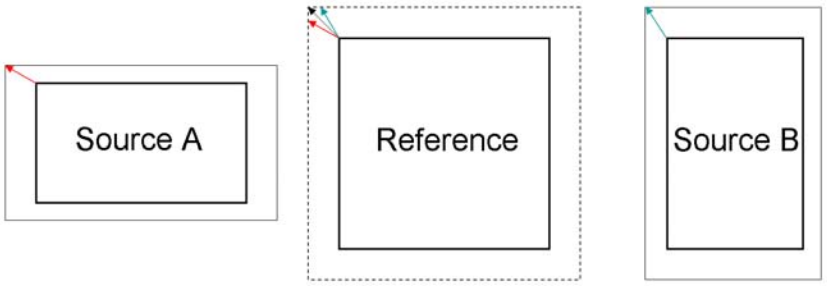


Figure 4.4: The finite strain reorientation [63], as used by Studholme and Cardenas [24] on Jacobian tensors, fails to reorient deformation vectors for the changes that occur in affine transformations that are not purely rigid, for example the anisotropic scaling shown here.

For the special case of affine inter-subject transformations, $T_{sr}(r) = Lr + t$ is spatially constant, as is $J_{sr} = L$, and we have $J_r(r_0) = L^{-1}J_s(s_0)L$. Studholme and Cardenas [24] almost have the corresponding expression $J_r(r_0) = SJ_s(s_0)S^{-1}$, with $S = L^{-1}$ except they have incorrectly replaced S^{-1} with S^T, which holds only if the linear transformation S is a rotation (without scaling or shearing).

Transformation of Jacobian determinant fields

The properties $|AB| = |A||B|$ (for square matrices) and $|A^{-1}| = 1/|A|$ (for invertible A), allow us to write

$$|J_r(r_0)| = |J_s(s_0)|\frac{|J_{sr}(r_0)|}{|J_{sr}(r_1)|}, \tag{4.14}$$

where we see (as Rao et al. showed) that it is not generally correct to simply resample the subject's longitudinal Jacobian determinant image J_s at the point $s_0 = T_{sr}(r_0)$ to find

the equivalent Jacobian determinant in the reference space at r_0; instead, a scaling factor depending on the change in the inter-subject Jacobian determinant J_{sr} between r_0 and r_1 is required.

For the special case of affine $T_{sr}(r) = Lr + t$, $|J_{sr}| = |L|$) is equal at r_0 and r_1, so we find that it is sufficient to resample the subject's Jacobian determinant image in the reference space. This contradicts Studholme and Cardenas [24], who wrongly state 'for a subject with a temporal lobe which is twice a big as another subject, their atrophy rate will be increased by a factor of two when mapping the change deformations into the reference space.' Rao et al. [67] make a further interesting observation that the behaviour of the longitudinal deformations is qualitatively the same in the local neighbourhoods of the fixed points $T_r(r_f) = r_f$ and $T_s(s_f) = s_f$, additionally showing that these fixed points correspond: $s_f = T_{sr}(r_f)$).

Interestingly, although Chung et al. [50] appear not to have considered the potential need for special methods of resampling longitudinal TBM data, we note here that the volume dilatation, which is equivalent to the trace of the Jacobian, is also invariant under equation (4.13) with an affine transformation: $\mathrm{tr}\left(J_r(r_0)\right) = \mathrm{tr}\left(L^{-1}J_s(s_0)L\right) = \mathrm{tr}\left(J_s(s_0)\right)$.[26] However, the importance of this result is very limited, since in practice, the inter-subject normalisation is unlikely to be affine, and in addition, the intra-subject deformation is likely to be large enough to make the determinant preferable to the dilatation (this is tested in the experimental results later).

Transformation of strain tensor fields

Returning to equation 4.13, we now consider SPD tensors derived from J. For the squared left and right stretch tensors, we have respectively

$$J_r(r_0)J_r^T(r_0) = J_{sr}^{-1}(r_1)J_s(s_0)J_{sr}(r_0)J_{sr}^T(r_0)J_s^T(s_0)J_{sr}^{-T}(r_1),$$
$$J_r(r_0)^T J_r(r_0) = J_{sr}^T(r_0)J_s^T(s_0)J_{sr}^{-T}(r_1)J_{sr}^{-1}(r_1)J_s(s_0)J_{sr}(r_0).$$

In the special case of a rigid transformation $J_{sr} = R$, $J_{sr}^{-1} = R^T$, and the above expressions reduce to

$$J_r(r_0)J_r^T(r_0) = R^T J_s(s_0)J_s^T(s_0)R,$$
$$J_r(r_0)^T J_r(r_0) = R^T J_s^T(s_0)J_s(s_0)R.$$

These expressions mirror that of Alexander et al. [63] for rotation of a diffusion tensor (though Alexander et al. have R as the mapping from source to reference, and hence R and $R^{-1} = R^T$ are exchanged).

At this point it is crucial to note that if T_{sr} is actually non-rigid, we do not wish to derive a rotation matrix from it, as is done by Alexander et al. [63] in the diffusion tensor case, due to the fundamental distinction between microscopic diffusion and macroscopic

[26] In fact, all three eigenvalues (which can be used to derive both trace and determinant) are preserved under the matrix similarity if the transformation is affine.

serial deformation, as explained by Rao et al. [67]. Our first pair of expressions above cannot therefore be simplified in general.

Diffusion tensor reorientation

It is of interest to consider diffusion tensor reorientation more closely, as there is a considerably more prior work than in the field of longitudinal morphometry. Figure 4.5 demonstrates the situation in diffusion imaging, where the desire is to preserve the properties of tissue microstructure while simultaneously preserving the macroscopic continuity of fibre tracts. In order to preserve the microstructural properties, the eigenvalues of the diffusion tensor must be preserved — if a brain is larger in a particular dimension, it is assumed that it has larger tracts, but the same diffusion coefficient along those tracts [63]. While maintaining the eigenvalues, the eigenvectors must be appropriately reoriented. One solution, which seems not to appear in the DTI literature, is to replace the eigenvectors with those from the transformed tensor – since affine transformations map ellipsoids to ellipsoids it is possible to rotate the original ellipsoidal axes into alignment with the new ones while preserving their original lengths or tensor eigenvalues. Since the major eigenvector is of greatest importance in subsequent tractography, Alexander et al. [63] instead proposed an approach to preserve this principal direction of diffusion (PPD): the reoriented tensor has its first principal eigenvector along the direction of the linearly transformed original tensor's first principal eigenvector. Since each direction corresponds to two degrees of freedom, while a rotation can only adjust three, it is not possible to orient the second principal eigenvector along the transformed second eigenvector direction if the transformation has skewed the axes.[27] The closest approximation to this, in terms of minimal angle between the result ($\tilde{v}_2$) and the transformed second eigenvector (Jv_2), is achieved when v_2 is rotated into the plane spanned by Jv_1 and Jv_2. Orthogonality then constrains the third eigenvector entirely.

We observe here, novelly to the best of our knowledge, that the new PPD eigenvectors are simply equal to the serially-orthonormalised transformed eigenvectors, i.e. Jv_1 normalised, Jv_2 orthogonalised with respect to Jv_1 and then normalised, and Jv_3 orthogonalised with respect to both Jv_1 and Jv_2 and then normalised. The rotation matrix can then be obtained by treating the orthonormal matrices of eigenvectors as rotation matrices and taking their ratio: $\tilde{V}V^{-1} = \tilde{V}V^T$, where $\tilde{V}$ collects together the new eigenvectors and V collects the original ones.[28]

[27] The transformation does not actually need to contain skews in order to do this, since anisotropic scaling alone will skew a set of axes rotated with respect to the axes of the scaling.

[28] Brief evaluation of this appears to show that our approach is computationally superior to the traditional use of Rodrigues' rotation formula to derive the two separate rotation matrices from angles and axes of rotation, as suggested in Alexander et al, and implemented in e.g. FSL's vec_reg (`http://www.fmrib.ox.ac.uk/fsl/fdt/fdt_vecreg.html`), however, further investigation of this is clearly outside the scope of the present thesis.

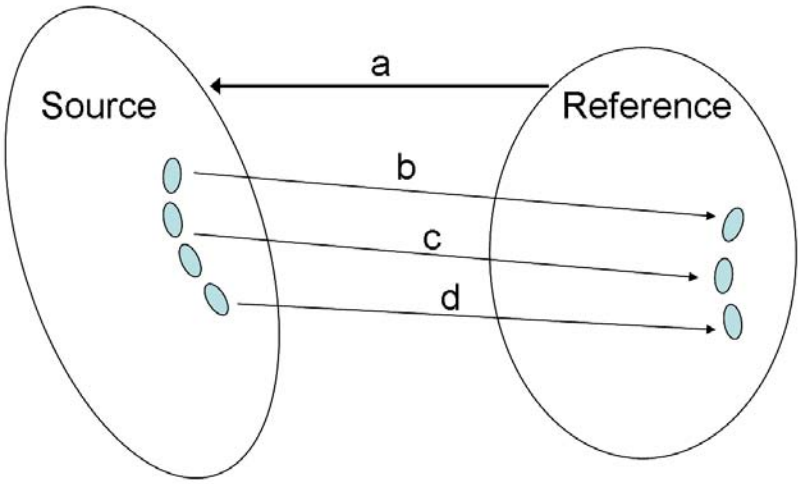


Figure 4.5: (a) Macroscopic transformation of anatomy, according to image registration. (b) Microscopic properties of water diffusion are preserved (eigenvalues of diffusion tensor, or size and shape of ellipse shown here). (c) Macroscopic compression is assumed to represent shorter tracts rather than shorter diffusion scale; hence the reduction from two to one equally-sized ellipses. (d) The orientation of each ellipse is transformed according to the anatomical transformation.

Procrustean diffusion tensor reorientation

We can now also make a novel connection between Alexander et al's PPD algorithm and more recent work from Xu et al. [68, 69]. Xu et al. propose the use of Procrustes analysis (see appendix C) to estimate a tensor reorientation from multiple neighbouring voxels. Their papers seem to suggest that their algorithm is fundamentally different to PPD. However, it is clear that if all three vectors from a single voxel are subjected to orthogonal Procrustes analysis (see section C.3) then the singular value decomposition of $\tilde{V}V^T = USW^T$, where $S = I$, and, trivially, $U = \tilde{V}$ and $W = V$ — meaning that the Procrustean solution of $R = UW^T$ is the same rotation obtained above for PPD. Interestingly, the Procrustes solution is also the same if the third vectors are dropped, in this case, the third singular value is zero, because the product of the rank-2 matrices $[\tilde{v}_1\ \tilde{v}_2]$ and $[v_1\ v_2]$ is also rank-2, nevertheless the third columns of U and W are constrained to be orthogonal to the first two, and unit-norm, due to the properties of the SVD (see section A.2) and hence R is again the same. Having made this connection, the extension of PPD to multiple voxels in the Procrustean framework is now completely straightforward. To summarise the method: the original first and second eigenvectors (at just the voxel in question, or at it and its n neighbours) form the $3 \times 2n$ matrix V; the normalised transformed set of n first eigenvectors form half of $\tilde{V}$; the second half of $\tilde{V}$ contains the n transformed second eigenvectors, orthogonalised with respect to the first half (so that the multiple second eigenvectors do not affect the preservation of the first principal direction, as would be computed from Procrustes analysis on just the first halves) and subsequently normalised; the singular value decomposition $USW^T = \tilde{V}V^T$ then gives the rotation as $R = UW^T$; the reoriented tensor is of course given by RDR^T.

Finite-strain diffusion tensor reorientation

The 'finite strain' (FS) reorientation strategy, discussed by Alexander et al. [63], is again based on Procrustes analysis, but is subtly different from the method of Xu et al. It simply takes the Procrustean solution to the problem of best approximating a linear transformation with a rotation, as outlined in section C.3.1. Where $USV^T = F$ is the singular value decomposition of the linear transformation (or Jacobian matrix), FS reorientation uses the rotation matrix $R = UV^T$. As noted by Alexander et al. [63] This strategy ignores the fact that differently oriented ellipses are differently affected by a given transformation (e.g. a predominantly vertical ellipse has its major axis rotated by a horizontal skew, while a predominantly horizontal ellipse would have its major axis preserved by the same transformation).

Thanks to our Procrustean interpretation of Alexander's PPD algorithm in the previous paragraph, it is easier to see that the FS method may be viewed as an approximation to PPD. If the transformation induces little or no non-rotational deformation, then the orthonormalisation of the second half of $\tilde{V}$ above has correspondingly little effect, meaning $\tilde{V} \approx FV$ and hence $USW^T = \tilde{V}V^T \approx FVV^T$, if V contains all three eigenvectors from a single voxel then it is orthogonal, giving $FVV^T = F$ and hence the SVD of $\tilde{V}V^T \approx F$ approximates the SVD of F employed in the FS method. For use within a tensor registration algorithm, the FS approximation is computationally preferable to PPD, as discussed by Zhang et al. [70]. However, for post-registration transformation of tensor images, there is no reason not to use the more accurate reorientation, and indeed, the final resampling used in [70] does revert to PPD reorientation after using FS within the registration.

Regularisation of strain tensor reorientation

Having noted above, firstly that Rao et al.'s conjugacy results are the correct approach for longitudinal TBM, while additionally discussing the extension of PPD to multi-voxel Procrustean methods, it is natural to ask whether Rao et al's method could also be extended to include information from neighbouring voxels to help regularise the reorientation procedure. It seems likely that the answer is yes: one could probably use the unconstrained solution for the best linear transformation described in section C.2. However, the details of this are left for future work, since we instead follow the simpler approach described in the next section.

The problem of inversion and an alternative approach

The approaches presented above for transforming longitudinal deformation fields, Jacobian tensors, strain tensors and determinants require knowledge of the inverse of the transformation used to normalise the serial data to the fixed target image or atlas. For example, in equation (4.13), the inverse is needed in order to find the correct point $r_1 = T_r(r_0) = T_{sr}^{-1}(T_s(T_{sr}(r_0)))$ for evaluation of the Jacobian matrix inverse $J_{sr}^{-1}(r_1)$. Rao et al. [67] derive a numerical line integral technique to approximate T_r (in terms of the displacement offset u_r) without directly computing T_{sr}^{-1}.

Interestingly, it has been suggested that a closely related problem arises in the case of diffusion tensor reorientation; Alexander et al. [63] stated that use of the first order Taylor series approximation of a nonlinear transformation means that the Jacobian of the transformation can straightforwardly take the place of the linear part of the affine transformation they focus upon. Xu et al. [69] later made the following argument: (i) resampling the tensors into the space of the target conventionally requires the transformation field to be defined over the space of the target, so that the data can be 'back-transformed' without leaving gaps in the transformed result; (ii) the reorientation itself, in contrast, requires the Jacobian that approximates the linear transformation which is applied to the tensors (at their original location). If just this forward transformation is computed, then the tensors can be reoriented using it and projected into the space of the target, but they will then need to be re-gridded using a non-uniform interpolation method, potentially leaving gaps in the result, as emphasised by Xu et al. [69] who refer to such gaps as 'seams'.

While this seems valid at first sight, and appears not to have been countered in the literature, we explain here why it is incorrect: while the inverse transformation is unknown at the tensor/voxel locations of the source, it is known at the non-integral voxel location in the source $s = T_{sr}(r)$ from which the tensor is resampled to place into the target-aligned result at (integer) location r, since this is simply the negation of the vector $u_{sr}(r)$ that goes from r to s. Moreover, noting that in DTI there is no longitudinal deformation to consider, equation (4.12) reduces to $J_{rs}(s) = J_{sr}^{-1}(r)$, showing clearly that the Jacobian of the inverse transformation at location s is simply given by the inverse of the known Jacobian matrix of the original transformation at location r. This is the approach used in Alexander et al.'s Camino 'Reorient' tool (`http://www.cs.ucl.ac.uk/research/medic/camino/index.htm`) and in FSL's `vec_reg` tool (`http://www.fmrib.ox.ac.uk/fsl/fdt/fdt_vecreg.html`).

In the case of morphometry though, the inverse transformation is needed at a different location from that to which the original transformation points, and the Jacobian here is unknown. One approach would be to assume that the intra-subject deformation is small enough to ignore, and that $J_{rs}(s_1) \approx J_{rs}(s_0) = J_{sr}^{-1}(r_0)$ which is known. Alternatively, one could assume that the inter-subject deformation is sufficiently small that the Jacobian of the inverse transformation (at the required point s_1) in the source image) is approximately equal to the inverse of the original transform's Jacobian matrix at the *same* (i.e. identical, rather than corresponding) point in the target, i.e. to assume:

$$\begin{aligned} \left.\frac{\partial T_{rs}(s)}{\partial s}\right|_{s=s_1} &= \left[\left.\frac{\partial T_{sr}(r)}{\partial r}\right|_{r=r_1}\right]^{-1} \\ &\approx \left[\left.\frac{\partial T_{sr}(r)}{\partial r}\right|_{r=s_1}\right]^{-1}. \end{aligned}$$

However, as argued by Ashburner [37], these 'small deformation' approximations can be very poor, either in the presence of large anatomical variability between the brains of different subjects (especially if abnormal or atrophied patients are included) or in the

presence of significant longitudinal change. Therefore, in longitudinal TBM, one could correctly argue, as Xu et al. apparently incorrectly did regarding DTI, that a true bijection is required. Xu et al. [69] proposed an algorithm to find the inverse of a deformation field, which is somewhat unclear, but appears to be a standard scattered data interpolation [71] of the target coordinates over the space of the source image from their non-uniformly spaced transformed locations. In contrast, Ashburner's DARTEL algorithm instead ensures inverse-consistency by construction, since forward and backward transformations are generated by exponentiating (integrating) a velocity field or its negation.[29] Strictly, DARTEL is only exactly inverse consistent as the number of integration steps tends to infinity. For a single integration time-step, it reduces to a small-deformation approximation; for a typical number of steps (64 is the default [74]), there will be a small but measurable discrepancy. Finally, as mentioned earlier, Rao et al. [67] build an approximation to the effect of the inverse transformation without directly computing it, with a numerical line integral employing the inverse Jacobian matrices.

None of the above methods for generating an inverse transformation or its effect will be precisely correct (in the sense that $T_{rs}^{-1}(T_{rs}(x)) \neq x$), and those that are the most accurate are also the most computationally involved. An additional source of imprecision is that the inverse is required over the space of the longitudinally deformed source image (from equation (4.10), $r_1 = T_{sr}^{-1}(s_1)$) while the transformation to be inverted was derived to map the reference to the undeformed source; if large expansions are present, extrapolating the reference-source mapping to positions in the expanded source may lead to low accuracy. This potential problem is likely to be limited here, since T_s is chosen to map from a baseline image to an atrophied (rather than expanded) repeat image, but it is still likely to introduce some additional inaccuracy in terms of the round-trip composition of transformations. Furthermore, even ignoring any errors in the inverse, resampling multivariate deformation fields (and/or Jacobians etc.) seems likely to introduce equivalent or greater interpolation errors compared to resampling scalar images.

For the above reasons, the following alternative option may be preferred: simply spatially normalising the time-series for each subject (after within-subject rigid or affine registration without interpolation) using a single inter-subject transformation,[30] and then computing the longitudinal deformation fields directly in standard space. This means that measures derived from the longitudinal deformation fields can be directly analysed, without further consideration of the inter-subject transformations. This has an additional minor computational advantage in that the longitudinal non-rigid registration can be estimated on isotropic data with a more appropriate field-of-view than may have been acquired, and without needing to account for the rigid component. This method should be similar in effect to Rao et al's approach, as it first creates an approximate real instantiation of the hypothetical time-1 reference image (shown dotted in fig. 4.3) and then directly estimates the transformation which is conjugate to the original, that Rao et al. derive theoretically

[29] Diffeomorphic registration algorithms frequently build the overall transformation from multiple compositions of small incremental transformations [37, 72, 73], for which the small-deformation approximation of an inverse is applicable.

[30] Derived, for example, from the baseline or the average of each time-series (as in section 3.3).

(d in fig. 4.3). However, further work would be required to experimentally validate this. Having mentioned the potential difficulty in extrapolating the reference-source mapping in the case of expansion of the source, we should admit here that the two-stage approach still suffers a similar problem, in that a significantly expanded source image could be poorly spatially-normalised by the transformation derived for its unexpanded original. Again, the choice of baseline images for the inter-subject registration should limit this problem for atrophic diseases (if brain growth is expected, for example in developmental studies, it seems preferable to estimate inter-subject warps from the latest time-point images instead).

4.3 Experimental methods

This section presents the results of applying the above methods to a particular study: a subset of the longitudinal MIRIAD data-set [8]. The subset comprises 36 probable Alzheimer's Disease patients and 20 age- and sex-matched controls. Standard T1-weighted 3D MRI were acquired at baseline, and at six- and twelve-month follow-up visits. Image acquisition followed the same protocol described in section 3.2.3.

4.3.1 Preprocessing

Following the discussion in section 4.2.10 it was decided to resample spatially normalised sets of longitudinal images first, rather than transforming initially-computed intra-subject deformation fields and/or Jacobians and/or strain tensors. The spatial normalisation was derived using the unified segmentation algorithm [75] in SPM5, applied to the baseline image of each subject. Subsequent time-points were then normalised with the same parameters as the baseline; all images were resampled using cubic B-spline interpolation [76] to avoid the blurring introduced by simpler trilinear interpolation.

Within each subject's spatially normalised set of longitudinal images, high-dimensional warping [77] was used to independently estimate registrations from 6- and 12-month repeat images to the baseline. From these deformation fields, the Jacobian matrix was estimated at every voxel using centred finite differencing; the other TBM measures are then simply derived from these matrices.

Smoothing, where necessary, was applied to each element of the (potentially) multivariate data separately, e.g. the x-, y-, and z-components of the deformation fields were independently smoothed with the chosen kernel.

4.3.2 Exploring smoothing

There is currently no practical objective way of deciding upon a 'correct' amount of smoothing to apply for high-resolution morphometric data. Bayesian methods developed in recent years (discussed briefly in section 4.5.2) can be used to automatically estimate the smoothness of signal in noise within each plane for fMRI, but they are currently too computationally demanding for three-dimensional estimation at reasonable resolutions.

Even if simulated atrophy data with known gold-standard volume changes was used,

the spatial scale of the underlying patterns of group-wise significant findings (after smoothing out residual inter-subject misregistration) would not be known.

An initial univariate TBM experiment is therefore performed to briefly investigate the need for smoothing, and once found desirable, to determine, approximately, the visually optimal choice of full-width at half-maximum (FWHM) for the isotropic Gaussian kernel. The investigation is carried out on the (scalar) Jacobian determinants, after log-transformation, as this type of measurement has been the most commonly analysed in TBM to date. The following smoothing kernels are compared: 0 (no smoothing), 4 mm, 8 mm and 12 mm FWHM. Kernels larger than 12 mm were found unworthy of investigation due to the visually apparent over-smoothing already present at 12 mm. A simple parametric two-group t-test is performed, with a one-sided alternative-hypothesis, testing for greater volumetric contraction in patients compared to controls. Results are presented for the contrast (t-numerator), unthresholded t-statistic, and t-statistic thresholded to control FWE at a level of 5%. Note that RFT correction is likely to be highly conservative without smoothing, in which case we follow SPM's default approach of falling back on Bonferroni FWE-correction if it is less stringent than RFT. Obviously, permutation-testing would be expected to be superior in this case, but this is of limited interest in this brief and largely qualitative exploration of smoothing options.

4.3.3 Statistical methods

In chapter 2 we develop theory for permutation-testing of fully general linear models, including categorical factors and nuisance covariates. We focus there on Wilks' Λ statistic, which derives from the generalised likelihood ratio test. In this chapter, we are solely interested in one of the simplest forms of design: the comparison of two independent samples (which are typically assumed to have common variance). Because the interest is in longitudinal change (over a relatively carefully controlled 12-month interval) with inter-subject spatial normalisation removing differences in overall brain volume, we avoid the need to covary for global measures such as total intracranial volume (TIV) or total parenchymal brain volume. In addition, we have recently shown [78] that adjusting for TIV largely removes the need to adjust for gender in VBM studies of healthy controls (there may still be disease-gender interactions, but these are often of interest, rather than being purely a confound). Since the groups in this study are also carefully matched [8], we choose to focus on a pure two-sample test, with no nuisance covariates, for which the permutation test is exact. It may appear that the testing procedure would also be straightforward in this case, but there are two complications which are dealt with in the following subsections, the first is related to high-dimensional measures (which is particularly crucial for the searchlight technique, as can be seen from table 4.3); and the second is related to the special case of comparing the orientations of principal strain directions (as suggested in section 4.2.9).

The problem of covariance matrix dimensionality

In the two-sample test, Wilks' Λ reduces to the special case of Hotelling's T^2 test (2.1). However, both of these statistics suffer from the 'curse of dimensionality' in that they rely on determinants of full $m \times m$ symmetric covariance matrices containing $m(m+1)/2$ unique elements. If the number of elements approaches the number of observations, their estimation will become unstable, and the value of Wilks' Λ correspondingly unreliable. Once there are more elements than observations, the covariance matrix becomes singular and Λ becomes undefined.

It is possible to estimate well-conditioned covariance matrices from inadequate data if one is willing to bias or 'shrink' the estimates in some way. Ledoit and Wolf [79] derived such an estimator which is the asymptotically optimal convex combination of the sample covariance matrix with a scaled identity matrix. In a classification setting, Thomaz et al. [80] derive an approach for combining singular class covariance matrices with a non-singular pooled covariance matrix using the maximum entropy principle. They later adapted this to combine a singular sample covariance matrix with a scaled identity in their 'maximum uncertainty' discriminant framework [81]. Essentially, while Ledoit and Wolf form a weighted average of sample and identity covariance matrices, Thomaz et al. form a new matrix by clipping the eigenvalues of the sample covariance matrix to be no lower than the average of all original eigenvalues. Thomaz et al. motivate this approach by objecting to the fact that shrinkage methods also unnecessarily adjust the larger eigenvalues, however, they have not rigorously defined under what assumptions and optimality criteria their estimate may be proven to be better.

For either of the above methods, the need to apply them for every permutation, and at every voxel, makes them computationally very costly. Kriegeskorte et al. [1] nevertheless used a shrinkage estimate in their work on the searchlight in fMRI. Here, the special simplicity of the design (together with the permutation-testing framework) means that other test statistics can be employed, which do not require covariance matrix estimation, and should perform better with large m. One such statistic is presented next.

The two-sample Cramér test

The Cramér test is a non-parametric test for comparing two samples of univariate or multivariate observations. It is based on the Euclidean interpoint distances between pairs of observations. Baringhaus and Franz [82] showed that for independent m-dimensional random vectors A_1, A_2, B_1, B_2, where A_1 and A_2 have the same distribution function F, and B_1 and B_2 have the same distribution function G,

$$\mathrm{E}\|A_1 - B_1\| - \tfrac{1}{2}\mathrm{E}\|A_1 - A_2\| - \tfrac{1}{2}\mathrm{E}\|B_1 - B_2\| \geq 0,$$

with equality holding if and only if F and G are the same. This motivates the definition of the Cramér statistic, which is the difference of the sum of all the Euclidean interpoint distances between two different samples and one-half the sum of the two corresponding sums of distances within the same sample [82]. Since the publication of the original paper,

Franz has extended his `cramer` software (`http://cran.r-project.org/web/packages/cramer/index.html`) to include the concept of a 'kernel' function ϕ, which modifies the Euclidean distance. Including this, with n multivariate observations in the rows of Y, of which n_1 are in group $\mathcal{G}_1$ and n_2 in group $\mathcal{G}_2$, the statistic is given by

$$t = \frac{n_1 n_2}{n_1 + n_2}\left[\frac{2}{n_1 n_2}\sum_{i\in\mathcal{G}_1}\sum_{j\in\mathcal{G}_2}\phi(\|y_i - y_j\|^2) - \frac{1}{n_1^2}\sum_{i\in\mathcal{G}_1}\sum_{j\in\mathcal{G}_1}\phi(\|y_i - y_j\|^2) - \frac{1}{n_2^2}\sum_{i\in\mathcal{G}_2}\sum_{j\in\mathcal{G}_2}\phi(\|y_i - y_j\|^2)\right] \tag{4.15}$$

The kernel function must be defined on the positive real line, with non-constant monotone first derivative, and $\phi(0) = 0$. Kernels available in the R software package are given in table 4.2, though we use only the original ϕ_{Cramer}.

Kernel function	Alternative hypothesis
$\phi_{Cramer}(z) = \sqrt{z}/2$	Location
$\phi_{log}(z) = \log(1+z)$	Location
$\phi_{FracA}(z) = 1 - 1/(1+z)$	Dispersion
$\phi_{FracA}(z) = 1 - 1/(1+z)^2$	Dispersion
$\phi_{Bahr}(z) = 1 - \exp(1 - z/2)$	Location or dispersion

Table 4.2: Cramér test kernels available in the R package `cramer`.

Baringhaus and Franz propose the use of resampling methods to determine critical values of the statistic [82], and their software provides three alternative approaches: conventional bootstrap, permutation-based resampling, and bootstrapping the limit distribution. Whitcher et al. [16] suggested the use of a non-stochastic procedure for approximating the asymptotic distribution based on the quadratic form of Normal random variables. Here, we use the permutation framework described in chapter 2 as a principled means to obtain both voxel-wise and family-wise critical thresholds (or p-values). Appendix D.4 considers the problem of computationally efficient implementation of permutation testing for the Cramér statistic.

The Bipolar Watson test

Section 4.2.9 presented the principal strain direction as a potentially important orientational tensor-based morphometry measure. We explained the need for a test which accounts for the Riemannian structure of the manifold of directions (axes), and mentioned the Bipolar Watson test developed for the related problem in diffusion tensor imaging by Schwartzman et al. [64]. Further details on this test will now be given.

The bipolar Watson distribution is one of the simplest distributions on the unit sphere which is antipodally symmetric [64], its (unnormalised) PDF is given by

$$f(\pm\hat{v}|\mu,\kappa) \propto \exp\left(\kappa(\mu^T\hat{v})^2\right).$$

The mean direction μ is also a unit vector, so $\mu^T\hat{v}$ is the cosine of the angle between the

axes $\hat{v}$ and μ. The positive constant κ is the 'concentration'; higher values result in more tightly clustered points around $\pm\mu$. Given a collection of observed unit vectors in the rows of the data matrix Y,[31] The maximum likelihood estimator of μ is given by the principal eigenvector u_1 of the scatter matrix of the observations —

$$S = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T = \frac{1}{n} Y^T Y.$$

The corresponding principal eigenvalue gives the 'sample dispersion' $s = 1 - \lambda_1(S)$, which becomes the maximum likelihood estimator of $1/\kappa$ asymptotically as $\kappa \to \infty$.

Schwartzman et al. [64] propose a decomposition of the total dispersion s in terms of the dispersions $s_i = 1 - \lambda_1(S_i)$ from the group scatter matrices $S_i = Y_i^T Y_i$, where Y_i is the matrix of observations in group i. Similar to an analysis of variance decomposition, for the two groups,[32] using

$$ns = (n_1 s_1 + n_2 s_2) + (ns - n_1 s_1 - n_2 s_2)$$

leads to a test-statistic

$$F = \frac{(n-2)(ns - n_1 s_1 - n_2 s_2)}{n_1 s_1 + n_2 s_2}. \tag{4.16}$$

This statistic is asymptotically F-distributed as $\kappa \to \infty$, which seems quite a strong assumption. We avoid this approximation by using the permutation-testing framework, for which it is important to make the calculation of the test statistic as efficient as possible. Writing

$$ns = n(1 - \lambda_1(S)) = n - \lambda_1(Y^T Y) = n - \lambda,$$

and

$$n_1 s_1 + n_2 s_2 = n - \lambda_1(S_1) - \lambda_1(S_2) = n - \gamma,$$

the following rearrangements of the test statistic are permutationally equivalent:

$$\begin{aligned} F \overset{p}{=} F_p &= \frac{ns - n_1 s_1 - n_2 s_2}{n_1 s_1 + n_2 s_2} \\ &\overset{p}{=} \frac{ns}{n_1 s_1 + n_2 s_2} \\ &\overset{p}{=} \frac{n - \lambda}{n - \gamma}. \end{aligned}$$

Now, for FWE inference (as we are primarily interested in here), λ varies over the voxels, so we cannot further simplify the test statistic $f = (n-\lambda)/(n-\gamma)$. However, for uncorrected (or FDR inference, as used in [64]) one could consider each voxel independently. $Y^T Y$ is invariant to permutation of the rows of Y, and therefore, we can treat $\lambda_1(Y^T Y) = \lambda$ as constant (for each voxel). Meaning that the statistic can be simplified further still:

$$F_p \overset{p}{=} F_v = \frac{1}{n - \gamma} \overset{p}{=} \gamma,$$

[31] We drop the notation $\hat{y}$ for a unit vector here, simply using y for conciseness.

[32] Straightforward extension to the k-sample problem is given in the appendix of [64].

where $\lambda \leq n$ can only achieve equality in the degenerate case that all axes are coincident.

This surprising new result has been confirmed with Monte Carlo evaluations: the original statistic in (4.16) and γ produce values under permutation of the data with the same sort-order. No further algebraic simplification is possible, since $\gamma = \lambda_1(S_1) + \lambda_1(S_2)$ involves a sum of the principal eigenvalues of two matrices which are different for each permutation. However, there remain opportunities for computational simplification (also relevant to our more general FWE-suited test statistic) in the implementation of the permutation-test, which we describe in appendix D.4.

Quantitative performance comparison

Rigorous evaluation of alternative neuroimaging methods is a challenging task (see e.g. [83, 84] and related comments in section 4.3.2). In chapter 3 we used simulated atrophy to derive gold-standard maps of expected change. However, this approach is far from perfect. Simulated deformation fields give ground-truth for within subject volume changes, but they do not give a straightforward gold-standard result for the pattern of group-wise differences that should be considered significant in light of intersubject variability and (with the low-dimensional DCT-normalisation used here) registration that only attempts to match large scale anatomical features. Further work in this area would be useful, however, in this chapter, we abandon the concept of simulated ground-truth and instead rely on the over-simplified approximation that 'more is better' in a comparison of healthy aging with Alzheimer's disease. Hua et al. [85] make essentially the same assumption, where they comment that:

> Although an approach that finds greater disease effect sizes is likely to be more accurate than one that fails to detect disease, it would be better to compare these models in a predictive design where ground truth regarding the dependent measure is known.

Though they offer no further details on how a realistically complicated scenario could have a known dependent measure.

With the above assumption, performance can be quantified by comparing p-values. A naïve approach would be to simply quote the most significant p-value found anywhere in the image under each of the methods/statistics. However, this is somewhat removed from the kind of outcome which is clinically of interest, since a single outlying significant voxel would not be considered as compelling as a more distributed (and not necessarily connected) pattern of less significant differences. A slight improvement might be to look at a more robust estimate of extremity, such as the 5^{th} percentile of the p-values, though this then ignores the actual degrees of significance for the voxels with more extreme p-values. A similar alternative, with similar limitations, would be to quote the numbers of supra-threshold voxels at an arbitrary significance level like 5%. The loss of information and arbitrariness of both these methods suggests a more complete graphical display of the counts of supra-threshold voxels as the significance level is varied from the most extreme p-value present through to unity. Methods with a few very strong voxels will then have

the highest counts at the strictest thresholds, while methods with large areas of moderate significance will move to the fore as the threshold becomes generous enough to include them. Expressing these counts in a slightly more generalisable way as fractions of the total number of voxels (which is of course also given by the number of supra-threshold voxels for a p-value of unity) transforms the curve into the cumulative distribution function (CDF) of the p-values. Hence, we use a p-value CDF measure very similar to that considered in section 2.5.1, except that the distribution is computed over the voxels of the image, instead of over the multiple Monte Carlo simulations. This kind of p-value CDF is intimately related to the FDR procedure [33], and has been used in the same context of method-evaluation by Lepore et al. [23] and Hua et al. [85].

Because the p-value CDF is computed over the voxels, it simply depends on the sorted vector of in-mask p-values — no information on the voxels' locations is employed. We therefore also use a complementary performance measure, which directly compares p-values from two or more methods at each voxel. In the absence of ground-truth at a particular voxel, it is difficult to interpret the difference in p-values across methods. We therefore propose to put these differences into context using their average, in a similar way to Bland-Altman plots [86]. In a comparison of two measures, their difference may be treated as independent to their sum (or average), and a plot of the difference against the average provides an informative summary. In order to generalise this to more than two measures, note that in the two-measure case, the difference between one of the measures and the mean,

$$p_2 - \bar{p} = p_2 - \frac{p_1 + p_2}{2} = \frac{p_2 - p_1}{2},$$

is simply a scaled version of the paired difference, and hence also satisfies the above independence. It is therefore natural to consider an extended Bland-Altman plot for multiple measures, taking the difference of each with respect to their overall mean. In the special case of two measures, the second difference is simply a negated version of the first, so only one need be plotted. A plot of differences against means for all voxels in the mask would be too cluttered to interpret though, so we must also consider how to summarise it. The approach taken here is to sort the voxels into order based on the mean p-value, and then to divide them into groups, before presenting summaries of the differences within each group. To be precise, we divide the mean p-values into an arbitrarily chosen number (8) of equally-sized groups,[33] covering a pre-specified range (from 0–0.1) potentially dropping some of the least significant p-values in order that the total number is a multiple of 8. We then summarise the voxel-averaged p-values by their 8 group-averages, and compute either averages or boxplots (showing median and interquartile range, as usual) of the voxel-wise differences in p-values within the bins. The following MATLAB code-fragment is provided to avoid any remaining ambiguity in this description:

```
pmax = 0.1; ngrp = 8;
p = p(mask, :); % M methods in columns
pm = mean(p, 2); pd = p - pm*ones(1, M);
```

[33] The desire to have equal numbers is the reason we use the sorting and grouping, instead of binning into fixed intervals.

```
nsub = nnz(pm <= pmax);
nper = floor(nsub / ngrp); nsub = ngrp * nper;
[pms inds] = sort(pm);
pms = reshape(pms(1:nsub), nper, ngrp);
pmm = mean(pms); % means of grouped values
pd = shiftdim(pd, 1); % size(pd) now [M nsub]
PD = zeros(M, nper, ngrp);
for m = 1:M
    PD(m, :, :) = reshape(pd(m, inds(1:nsub)), nper, ngrp);
end
pdm = squeeze(mean(PD, 2));
plot(pmm,zeros(size(pmm)),'k', pmm, pdm, 'x');
```

Boxplots of `squeeze(PD(m, :, :))` may also be produced for each method. We present both or whichever seems more informative in each case.

Our permutation-testing framework makes available three levels of corrected p-values: uncorrected, FDR, and FWE. We argue that FWE p-values are the most important for localisation of effects [87], and also wish to focus on these because they have been neglected in recent methodological work (e.g. [23, 64]). However, note that FWE p-values introduce a dependence on the smoothness of the underlying data that may complicate their interpretation. This is obviously true for Random Field Theory based FWE inference, which directly uses the estimated smoothness of the random field [29], but is also true for permutation-testing, which is implicitly affected by smoothness of the residuals via the distribution of the maximum-statistic (rougher residuals produce more extreme maxima by chance alone, hence the multiple-comparison correction is more severe, exactly as with RFT). This dependence on smoothness can be characterised more precisely in terms of the 'effective number of independent tests' — a concept explored in greater detail by Nichols and Hayasaka [87], and one which we would like to consider further for this application in future work. Unfortunately, in the present case it is only computationally feasible to record the permutation distribution of the maximum statistic; whereas the complete permutation distributions of each voxel is required to explore the effective number of independent tests. For this reason, uncorrected p-values (which are individually valid for comparison of different methods at corresponding voxels, regardless of the multiple-testing issue) are also essential to our performance evaluation. We avoid direct comparison of FDR p-values here, since their signal-adaptive property [33] (potentially permitting more significant adjusted p-values at a particular voxel thanks to the presence of significance elsewhere in the image) is unhelpful for method comparison [Thomas Nichols, private communication].

4.3.4 Deformation-based morphometry

Deformation-based morphometry is performed using 'mass-multivariate' analysis of the components of the displacement field at each voxel. Recall that this is in contrast to the definition of DBM in [2] as an overall multivariate 'characterisation of the differences in the vector fields that describe global or gross differences'. The displacement fields for the

baseline to 12-month follow-up longitudinal interval are analysed. The 0–6 month data were also studied, but are not reported as they exhibited qualitatively similar (though quantitatively weaker) patterns.

To shed further light on the need for smoothing, both raw unsmoothed images and data smoothed with an 8 mm FWHM Gaussian kernel are analysed. The modest multiplicity of the displacement vector fields (three components, or six unique covariance matrix elements) makes this an ideal setting to compare the two multivariate testing options, so all the analyses are repeated once with Wilks' Λ and once with the Cramér statistic.

4.3.5 Searchlight morphometry

The unsmoothed 0–12 month displacement fields are analysed using the searchlight technique with the Cramér test. Properties of a range of searchlight kernels are given in table 4.3. Only kernels of 4, 7, and 10 voxel squared radius are employed for the displacement fields. This choice was based on post-hoc observation of limited difference between results at the extremes of the range.

(a) r^2	1	2	3	4	5	6, 7	8	9	10
(b) width	1	1	1	2	2	2	2	3	3
(c) Nvox	7	19	27	33	57	81	93	123	147
(d) Ncov	28	190	378	561	1653	3321	4371	7626	10878
(e) Ncov3	231	1653	3321	4950	14706	29646	39060	68265	97461

Table 4.3: Spherical searchlight kernels. (a) squared radius in voxels, (b) half-width along axes (border required between blocks, see appendix D.3) (c) total voxels contained, (d) unique covariance matrix elements for scalar data, (e) unique covariances for displacement 3-vectors. Note that squared radii of 6 and 7 give equivalent searchlight kernels.

Searchlight tensor-based morphometry is also briefly investigated, using the commonly analysed log-transformed Jacobian determinants, with the multivariate searchlight kernel removing the need for the usual smoothing. Only searchlight kernels of more than 3 voxel squared radius are employed, given the need for a certain amount of smoothing (see sections 4.3.2 and 4.4.1). As proposed earlier, we additionally explore the use of spline-pyramid downsampled images, testing two levels of coarsened data with a searchlight kernel of squared radius 4 voxels.

4.3.6 Generalised Tensor-based morphometry

We investigate a range of multivariate and univariate Jacobian-derived TBM measures, listed in table 4.4. All the methods are tested on the 0–12 month data; the most promising are then applied over the more challenging 6 month interval. Note that we include the full Jacobian matrix, despite the theoretical limitations outlined at the end of section 4.2.6, because in practice, we found no negative determinants arose for our data.

The Cramér statistic is used for all these tests; selected data are then reanalysed with Wilks Λ for comparison in the section on methodological subtleties below.

The results focus on the question of whether multivariate generalised TBM measures are superior to the simpler univariate (log) determinant measure. We also compare the two univariate options of log determinant and trace of the Jacobian.

m	$m(m+1)/2$	Measure
1	1	Log-transformed determinant of Jacobian matrix
1	1	Trace of Jacobian matrix (or $\nabla \cdot u$)
1	1	Maximum eigenvalue of Hencky strain tensor
3	6	Eigenvalues of Hencky strain tensor
6	21	Hencky strain tensor elements
9	45	All elements of the Jacobian matrix

Table 4.4: Jacobian-derived TBM measures, listed by their dimensionality (m) and the corresponding number of unique covariance matrix elements.

Methodological subtleties

A number of aspects in the above TBM studies are open to further investigation; here, we explore some of the more subtle methodological issues, which have often been ignored in the literature. In particular, we thoroughly compare the interaction between the options for smoothing and for scalar or matrix log-transforming data in univariate and multivariate strain-tensor based morphometry.

Univariate TBM using the determinant of the Jacobian has been performed both with [19, 88] and without [26] the log-transformation. Several authors [89, 90] have argued that the logarithm is preferable (or even necessary) based on either principles of symmetry and inversion invariance or on statistical grounds. In multivariate or generalised TBM, essentially the same question arises, but with the matrix logarithm replacing the usual scalar (natural) logarithm. Recent work in generalised TBM [22, 23] has exclusively employed the matrix logarithm, on the justification provided by the elegant log-Euclidean framework accounting for the Riemannian nature of the strain tensor (see section 4.2.6). Earlier work [2], from the perspective of solid mechanics investigated different strain tensors with and without the matrix logarithm in their definition (see section 4.2.5 and table 4.1). In short, the Hencky tensor is $\operatorname{logm}(U)$, while analysis of the right stretch tensor U itself is equivalent to analysis of the Biot strain tensor, given by $U - I$, since the subtraction of a constant identity matrix has no impact on either the Cramér or Wilks' Λ tests.

In the context of diffusion weighted image analysis, Whitcher et al. [16] compared Cramér tests of the diffusion tensor with and without the matrix logarithm. Interestingly, they found that the log-Euclidean approach actually lowered the statistical significance of their findings, which motivates us to check whether this phenomenon is replicated in our morphometric data. If one moves away from the more mathematical arguments [58, 59, 91], to interpret either the scalar or matrix logarithm simply as a preprocessing step, then not only can it be viewed as optional, but one may also ask where in the preprocessing pipeline it should take place. Furthermore, one could employ the log-Euclidean framework as a means of smoothing the data, before returning the results to their original space

using either the scalar or matrix exponential as appropriate. While this might seem mathematically unjustified, it is compatible with a comment made by Whitcher et al. [16]:

> Whereas Riemannian metrics are advantageous in applications such as interpolation or regularization, from a hypothesis testing perspective there is no reason at this time to prefer the log-Euclidean distance.

In summary, one could consider any of the four schemes enumerated in table 4.5 for either the determinant or the right stretch tensor. Note that in terms of software implementation, all of the schemes are special cases of a three-stage process-smooth-process pipeline.

	Procedure			Notes
1		Smooth		Common in univariate TBM
2	Log	Smooth		Uncommon (possibly unused)
3		Smooth	Log	Most common in univariate and generalised TBM
4	Exp	Smooth	Log	Log-Euclidean smoothing only

Table 4.5: Smoothing options for strain tensors or determinants, including appropriate (scalar or matrix) exponential- and logarithm-transformations.

For the scalar determinant, there are two additional options of smoothing the Jacobian tensor itself before taking the determinant and optionally the logarithm. The strain tensor U could similarly be derived from a smoothed Jacobian, and, furthermore, the eigenvalues (and/or maximum eigenvalue) included in table 4.4 have a large number of options. However, in the interest of brevity, we neglect to compare these options, on the basis that they are not only less theoretically motivated, but also that they should approximately follow the behaviour for the scalar case. For example, if the scalar situation shows that smoothing the Jacobian matrix directly is inferior to all the other options, then it would be surprising to find that this was the best of the available options for the strain tensor or its eigenvalues.

It might initially seem that the volume dilatation $1 + \operatorname{tr}(K)$, or the statistically equivalent transformation divergence $\operatorname{tr}(J)$, present similar options. However, note that while $1 + \operatorname{tr}(K)$ was shown in section 4.2.4 and figure 4.1 to approximate $|J|$, there is no guarantee that $1 + \operatorname{tr}(K)$ or even $\operatorname{tr}(J) = 3 + \operatorname{tr}(K)$ will be positive, so the log-transformation is not generally applicable. The comparison of the volume dilatation to the determinant is still of interest, but this is addressed in the main TBM results section, comparing the trace of the Jacobian to the particular log-transformed case of the determinant.

Two further tensor-related subtleties that have been explained in section 4.2 are evaluated in practice here. Firstly, the infinitesimal strain tensor $(J + J^T)/2$ (see section 4.2.5) is compared to the finite strain Hencky tensor $\operatorname{logm}\left(\left(J^T J\right)^{1/2}\right)$, in each case taking the unique elements of these symmetric matrices. This comparison seems to be the closest multivariate analogue of the comparison between the dilatation and the log-determinant, though one could also argue that the infinitesimal strain tensor should be compared to the Biot or Green strain tensors from table 4.1. Secondly, we explore the distinction reached at the end of section 4.2.7 between analysing the elements of the Hencky tensor $\operatorname{vech}(H)$,

compared to true log-Euclidean analysis (yielding equal Frobenius and vector norms) of $\text{vech}_{LE}(H)$ with off-diagonal elements scaled by $\sqrt{2}$.

Another issue investigated in this section is the relative performance of the more general Wilks' Λ statistic compared to the design-specific Cramér test. Specifically, one measure of each dimensionality in table 4.4 is tested using both statistics; the commonly studied log-determinant being chosen to represent the univariate measures.

Finally, we briefly evaluate the benefit from the step-down procedure [92] for deriving FWE-corrected p-values, in comparison to the standard approach using the permutation distribution of the image-wise maximum-statistic [93].

Orientational measures

The orientational or directional measures discussed in section 4.2.9 and summarised in table 4.6 are explored in some detail. Firstly, their raw data are visualised for an individual AD patient. Secondly the group-wise arithmetic means over the 36 AD patients and 20 matched controls are illustrated. Finally, statistical results are presented. Tests are performed using the Cramér statistic, except for the unscaled principal direction, which is tested using the bipolar Watson distribution, as described in section 4.3.3.

m	$m(m+1)/2$	Measure
1	1	Geodesic anisotropy
3(2)	6(3)	Principal eigenvector direction
3	6	Principal eigenvector scaled by its (max) eigenvalue
3	6	Infinitesimal rotation tensor elements ($\nabla \times u$)

Table 4.6: Measures of anisotropy, orientation, or vorticity, derived from the Jacobian, listed by their dimensionality (m) and the corresponding number of unique covariance matrix elements. Numbers in parenthesis indicate a smaller number of true degrees of freedom than the dimensionality (since unit vectors can be identified with points on the surface of the unit-sphere).

Cross-methodological comparisons

Concluding this section of experimental work on morphometry, we compare the results from deformation-based morphometry (using the three-vector of displacement field components) to some representative results from tensor-based morphometry: the near-standard log-determinant, the equal-dimensional three-vector of eigenvalues of the Hencky tensor, the six unique elements of H, and the full Jacobian matrix. In addition, we include the geodesic anisotropy in this comparison, as the most powerful of the orientational measures (see sections 4.3.6 and 4.4.4).

4.4 Results and discussion

4.4.1 Smoothness comparison

Results for univariate TBM on the log-determinant of the Jacobian at the four different levels of smoothing are illustrated in the following figures. Figure 4.6 shows the t-statistic and the signal or effect-size on which it is based — i.e. the numerator of the t-statistic or 'contrast'. Figure 4.7 shows maximum intensity projections for the significant (5% family-wise error corrected) voxels at each level of smoothing.

Even with no smoothing, the contrast image shows a biologically plausible pattern of losses predominantly in the temporal lobe and insula, though the map is very noisy, and contains some more questionable isolated peaks, for example adjacent to the ventricles. However, with no smoothing, virtually nothing survives correction for multiple comparisons. With 4 mm, the contrast and t images show quite appealing patterns, but there is still substantial noise. The stringently corrected FWE results remove much of the spurious noise, leaving findings in anatomically reasonable locations; however, these are unrealistically speckled and noise-like. At higher levels of smoothing, there is a general tendency for spatially consistent findings to congeal into more interpretable contiguous regions of significance, however, at 12 mm (or above) there is some evidence that genuine anatomical detail has been lost due to excessive smoothing.

Based on subjective visual assessment, we favour the 8 mm FWHM results, which mirrors the findings of Scahill [94], who favoured the same amount for voxel-based morphometry. All further smoothed (i.e. not using the searchlight) results in this chapter are performed with 8 mm FWHM.

It must be admitted that we have severely under-sampled scale-space [34], in the sense that the changes from 4 to 8 and 8 to 12 mm are relatively pronounced; visual comparison of e.g. 7, 8, and 9mm may have indicated a slight preference for one of the other kernels. In addition, as mentioned briefly by Jones et al. [95], a logarithmic sampling of scale-space (e.g. 2, 4, 8, 16mm) would be more appropriate than our equally spaced samples. However, given the essential subjectivity of any preferences, and the potential inaccuracy when extrapolating from these univariate results to the various multivariate measures, these issues do not seem worth pursuing here. A more promising line of research — objective and automatic estimation of the signal smoothness using a principled Bayesian formulation — is very briefly discussed in section 4.5.2.

4.4.2 Deformation-based morphometry

Statistical results from the analyses of the baseline to 12-month follow-up displacement fields are summarised in figure 4.8. The first column shows the raw statistic maps for Wilks' Λ (actually its reciprocal, as stated in section 2.3.2, which is larger for greater differences) and the Cramér test (also larger for greater significance, but starting from 0 instead of 1 as Λ does). The figure's second and third columns illustrate the multiple-comparison corrected significant findings, using both FDR and FWE p-values, displayed after (negated) logarithmic transformation so that more significant voxels appear brighter.

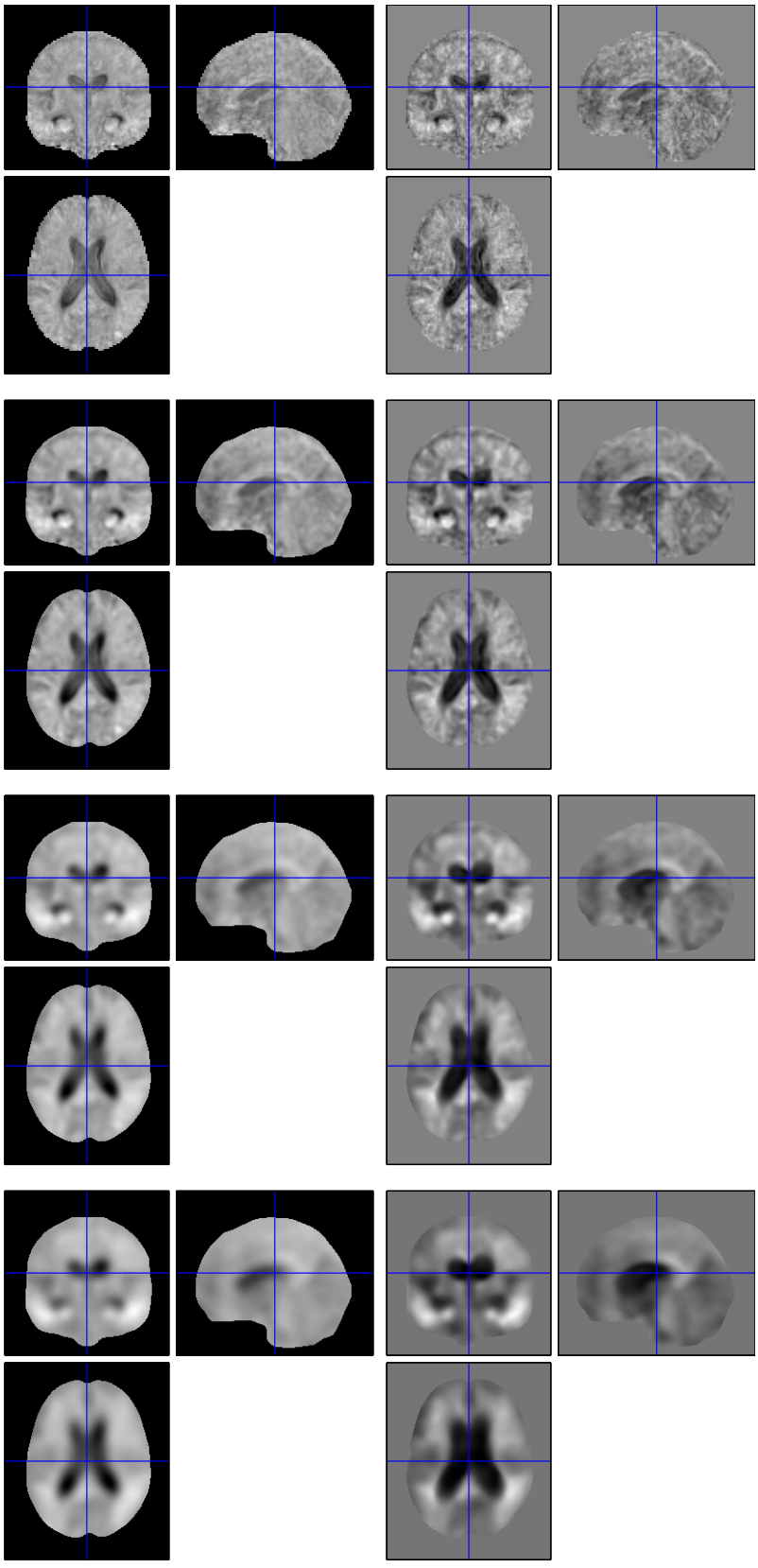
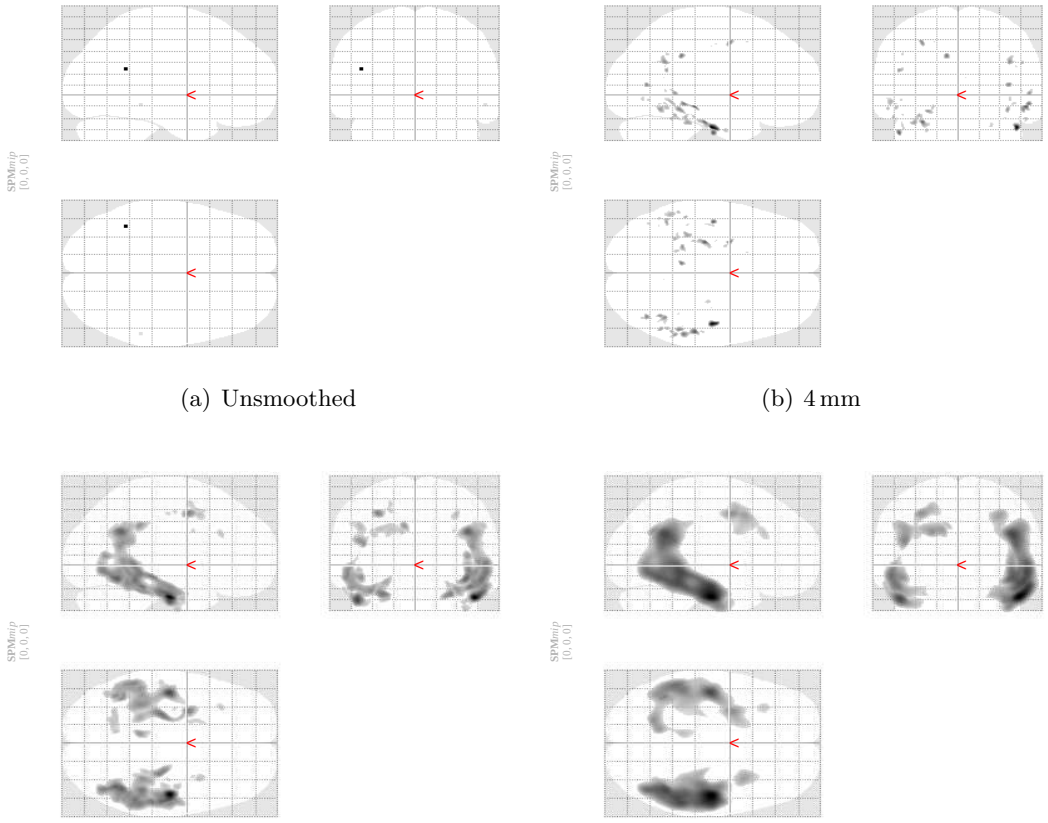


Figure 4.6: Results at different levels of smoothing. From top to bottom: 0, 4, 8 and 12 mm FWHM Gaussian kernel. Left column, contrast numerator; right column, t-statistic. Anatomical-left corresponds to display-left; cross-hairs are shown at the centre voxel.



(a) Unsmoothed (b) 4 mm

(c) 8 mm (d) 12 mm

Figure 4.7: Maximum intensity projections of significant findings at each level of smoothing. The t-statistics are thresholded to control $pFWE < 0.05$. Anatomical-left corresponds to display-left.

In terms of the FDR results, there is little to choose between the methods. Smoothing marginally increases power, while slightly reducing spatial acuity, but, interestingly, the difference in power is very small compared to the FWE-corrected results. The significant ($pFWE < 0.05$) areas are additionally illustrated in figure 4.9 as Maximum Intensity Projections (MIPs). As one might have expected, the permutation-testing FWE-correction based on the distribution of the image-wise maximum statistic, is less weakened by roughness than the parametric RFT results presented in section 4.4.1; nevertheless, there is a clear gain in power for both Wilks' Λ and the Cramér test obtained by smoothing.

The effect of smoothing and choice of statistical test are further visualised in figure 4.10. In this particular study, there seems to be little evidence that smoothing is degrading the accuracy of the spatial locations (cf. [96]); instead it appears to be increasing the extent of findings without overly displacing their peak locations. For either raw or smoothed data, the Cramér test appears to have dramatically greater power than the conventional Wilks' Λ when considering the FWE-corrected results.

Figure 4.11(a) more quantitatively investigates power, by showing cumulative distribution functions for the uncorrected p-values. The Cramér statistic appears superior for all thresholds stricter than about 0.01 (which is very lenient for uncorrected results), with

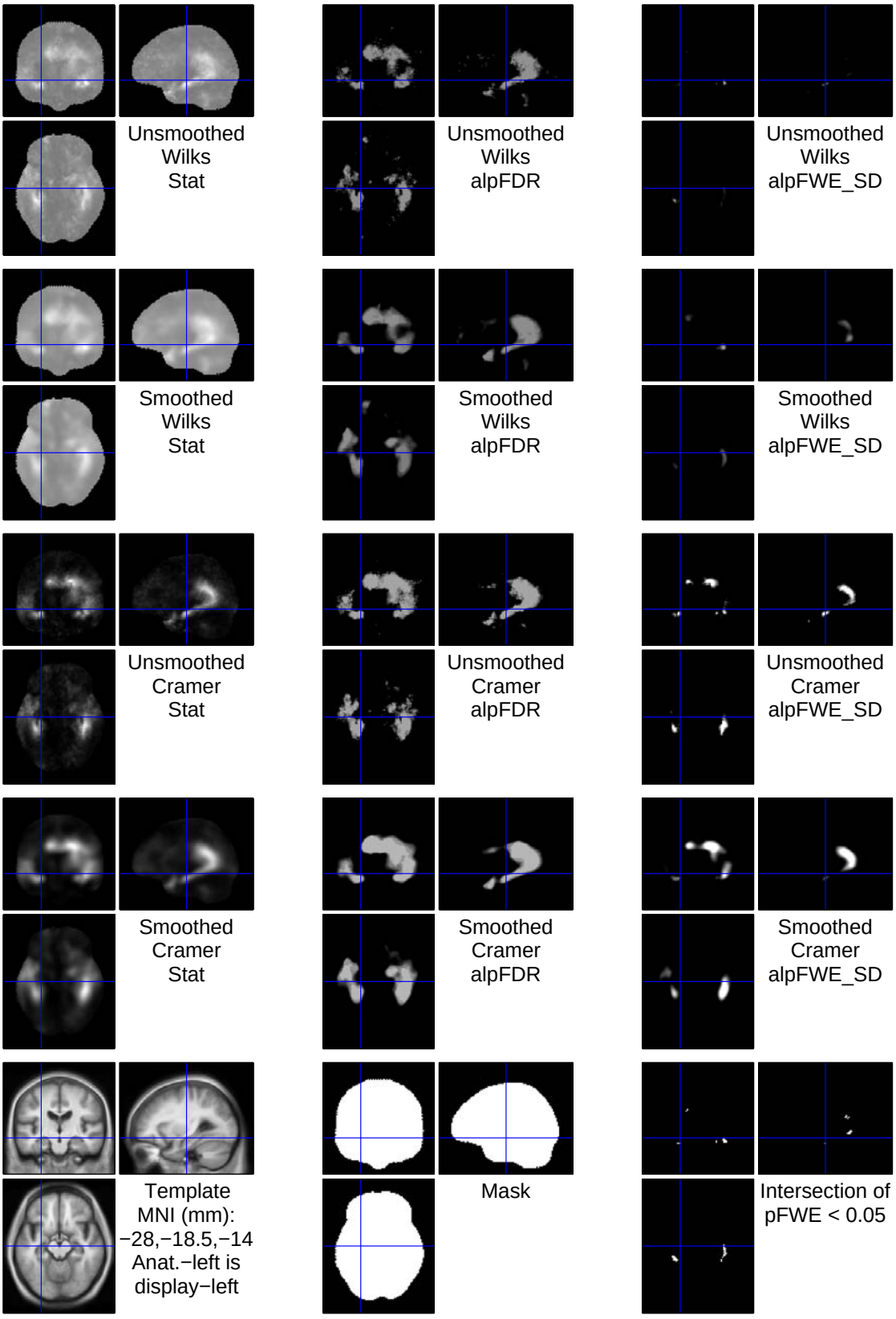
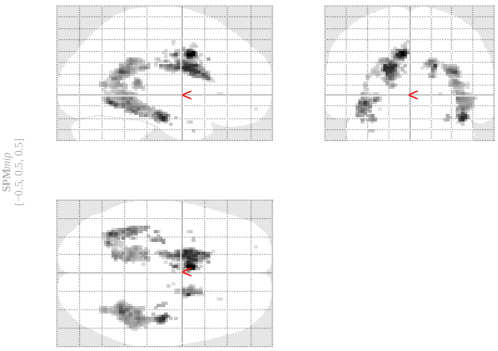
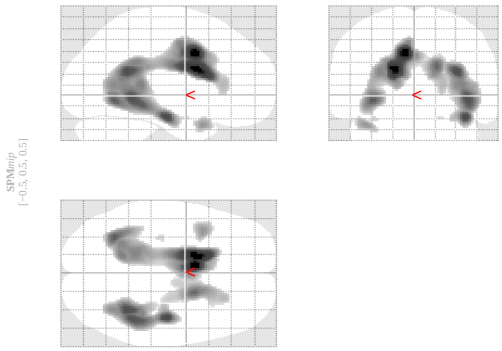


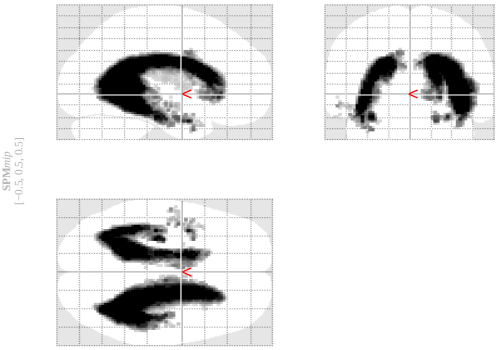
Figure 4.8: Statistical results for deformation-based morphometry, using Wilks' Λ and Cramér tests, on raw and 8 mm-smoothed displacement fields. P-values are displayed in the range 0.05–0.0005 as absolute log10 p-values (brighter is more significant). The final row shows the template, mask, and a Boolean intersection of the significant results of the first four rows, to provide context.



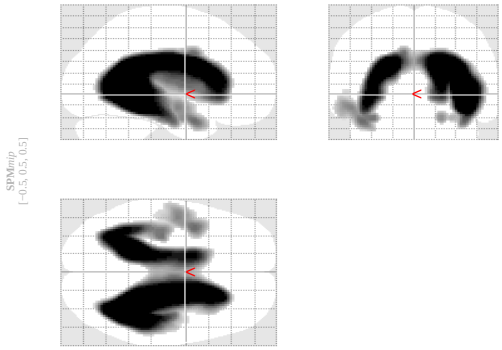
(a) Unsmoothed Wilks



(b) Smoothed Wilks



(c) Unsmoothed Cramer



(d) Smoothed Cramer

Figure 4.9: Maximum intensity projections of significant findings for DBM using Wilks' Λ and Cramér tests, on raw and 8 mm-smoothed displacement fields. Absolute log10 p-values over the range $0.00005 < pFWE < 0.05$ are shown. Anatomical-left corresponds to display-left.

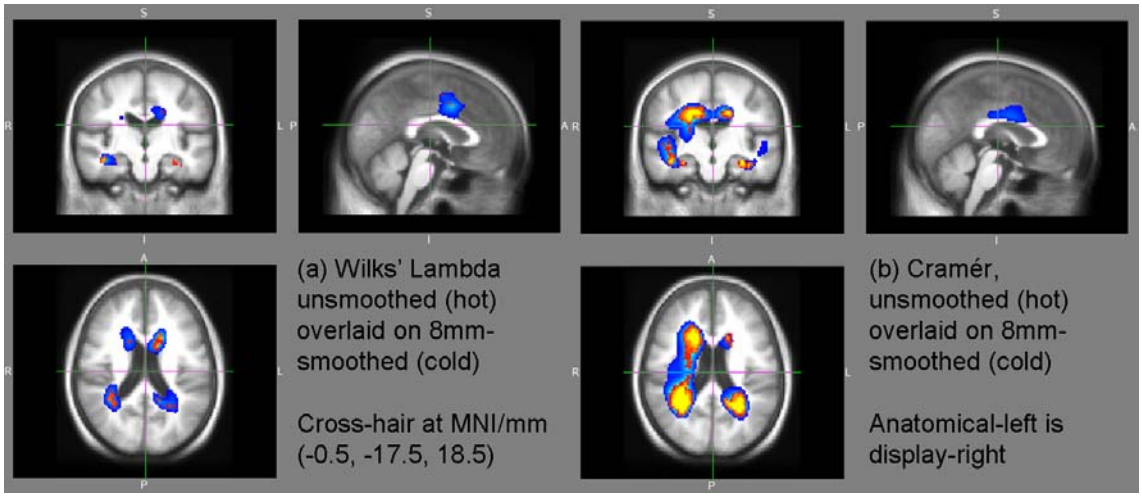
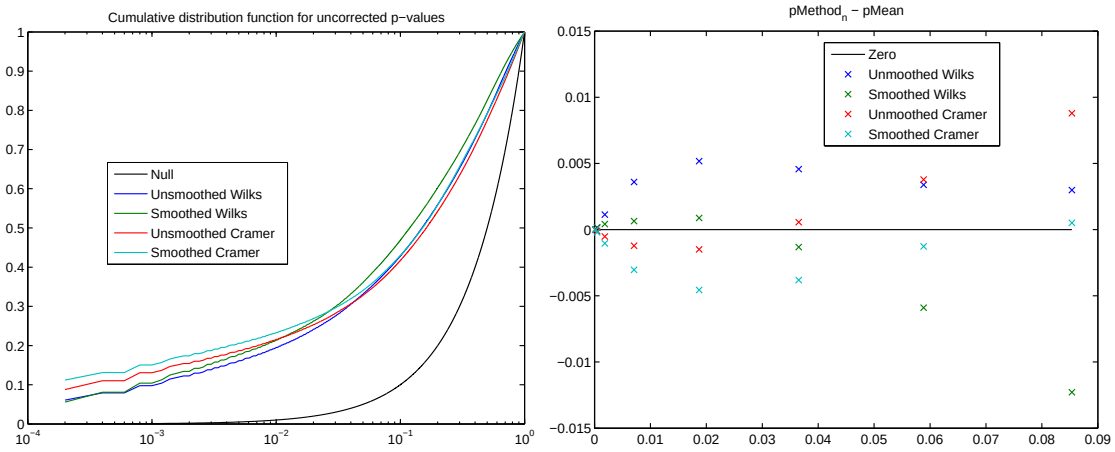


Figure 4.10: Overlays of significant ($0.0005 < pFWE < 0.05$) absolute log10 p-values for DBM using Wilks' Λ and Cramér tests, on raw and 8 mm-smoothed displacement fields.

the 8 mm FWHM Gaussian smoothing increasing the power for both statistics. CDF plots based on corrected p-values are not presented, but show essentially the same results, with the smoothed Cramér statistic performing best, followed by its unsmoothed version, and then the two Wilks' Λ tests. Panel (b) provides a comparison of the power of the different methods when evaluated over corresponding voxels, and similarly favours the smoothed Cramér statistic at reasonably strict thresholds. Wilks' Λ becomes more powerful at very relaxed levels of significance, and again, the smoothed data uniformly outperform the raw data using either statistic.



(a) pUnc CDFs (b) pUnc comp.

Figure 4.11: Statistical power of the different DBM tests illustrated via (a) cumulative distribution functions and (b) voxel-matched comparison of their uncorrected (permutation-based) p-values.

4.4.3 Searchlight morphometry

Continuing the analysis of displacement fields from the previous section, we now turn from using simple smoothing of the data prior to statistical analysis, to instead analysing the unsmoothed data within differently sized searchlight kernels. The Cramér statistic is used exclusively here, partly because of its superior power reported in the previous section, but mainly because the higher-dimensional multivariate measurements obtained by collecting three-vectors from all voxels in the searchlight would make the necessary covariance matrix estimation for Wilks' Λ particularly challenging.

Figure 4.12 summarises the findings, recapitulating the results from figure 4.8 for the unsmoothed and 8 mm-smoothed Cramér test *without searchlight* for comparison. The statistical maps show the expected gradual increase in smoothness with increasing searchlight radius; the largest 10 voxel squared-radius kernel appears to have produced a visually similar level of smoothing to the 8 mm Gaussian smoothing kernel. Approximate quantitative estimates of the smoothness of the statistic images are given in table 4.7, estimated with the `3dFWHMx` program from the AFNI software package.[34] This program estimates the smoothness along each dimension based on the sample variance of the numerical first

[34] Analysis of Functional NeuroImaging `http://afni.nimh.nih.gov/afni`

derivatives, using a three-dimensional extension of the expression derived in the appendix of Forman et al. [97]. The table shows a peculiar tendency for the anterior-posterior (y) smoothness to be higher than the others, which is almost certainly an artefact (perhaps reflecting a 'preferred direction' of the ventricles and/or hippocampal changes that dominate the statistic images). The geometric mean smoothness places the 8 mm smoothing slightly beyond the largest searchlight kernel considered, in rough agreement with visual inspection.

Searchlight r^2	FWHM_x	FWHM_y	FWHM_z	$\sqrt[3]{\prod_i \text{FWHM}_i}$
0	9.38	13.05	10.48	10.87
4	10.59	14.89	11.82	12.31
7	11.27	16.14	12.66	13.20
10	11.63	16.68	13.08	13.64
8 mm smoothing	12.32	17.81	13.92	14.51

Table 4.7: Smoothness of searchlight DBM statistic images in terms of Gaussian Full-Width at Half-Maximum, in mm, estimated using AFNI's `3dFWHMx`. The special cases of no smoothing and no searchlight ($r^2 = 0$) and of conventional smoothing, are included for comparison. The final column gives the geometric mean of the three directional smoothness values, because this corresponds to the scaled-identity covariance matrix closest, in the Riemannian sense, to the anisotropic diagonal covariance matrix (compare equation (4.8) for the geodesic anisotropy).

Considering now the p-value maps in figure 4.12, there is surprisingly little difference between the alternatives, either in terms of FDR or FWE significance, other than the fact that any searchlight kernel or smoothing is more powerful than the unsmoothed data analysed directly. There are perhaps slightly more voxels satisfying $pFWE < 0.05$ for the $r^2 = 10$ searchlight than for any of the other options, but this is naturally accompanied by slightly more blurring than the smaller kernels.

Observed powers are compared quantitatively in figure 4.13, which presents cumulative distribution functions and matched-voxel comparisons, using both uncorrected and FWE p-values. There is generally good agreement between the two correction methods, particularly in the matched-voxel comparisons which both show all three searchlight kernels to be superior to smoothing. The uncorrected CDFs favour searchlight with kernels of $r^2 = 7$ or 10 to classical smoothing; the FWE results show the $r^2 = 7$ kernel and the 8 mm smoothing performing approximately equally well, and slightly better than the smaller kernel. The largest kernel seems to have relatively poor FWE-corrected power at the more stringent levels, but has the highest CDF above about 1%. In the matched-voxel comparisons there is a tendency for the larger kernels to be the best at significance levels more lenient than 0.01, while the 33-voxel kernel ($r^2 = 4$) appears better for the very strictest alpha.

Searchlight TBM

We now test the searchlight technique on the most common univariate tensor-based morphometry measure, the log-transformed determinant of the Jacobian, based on the same

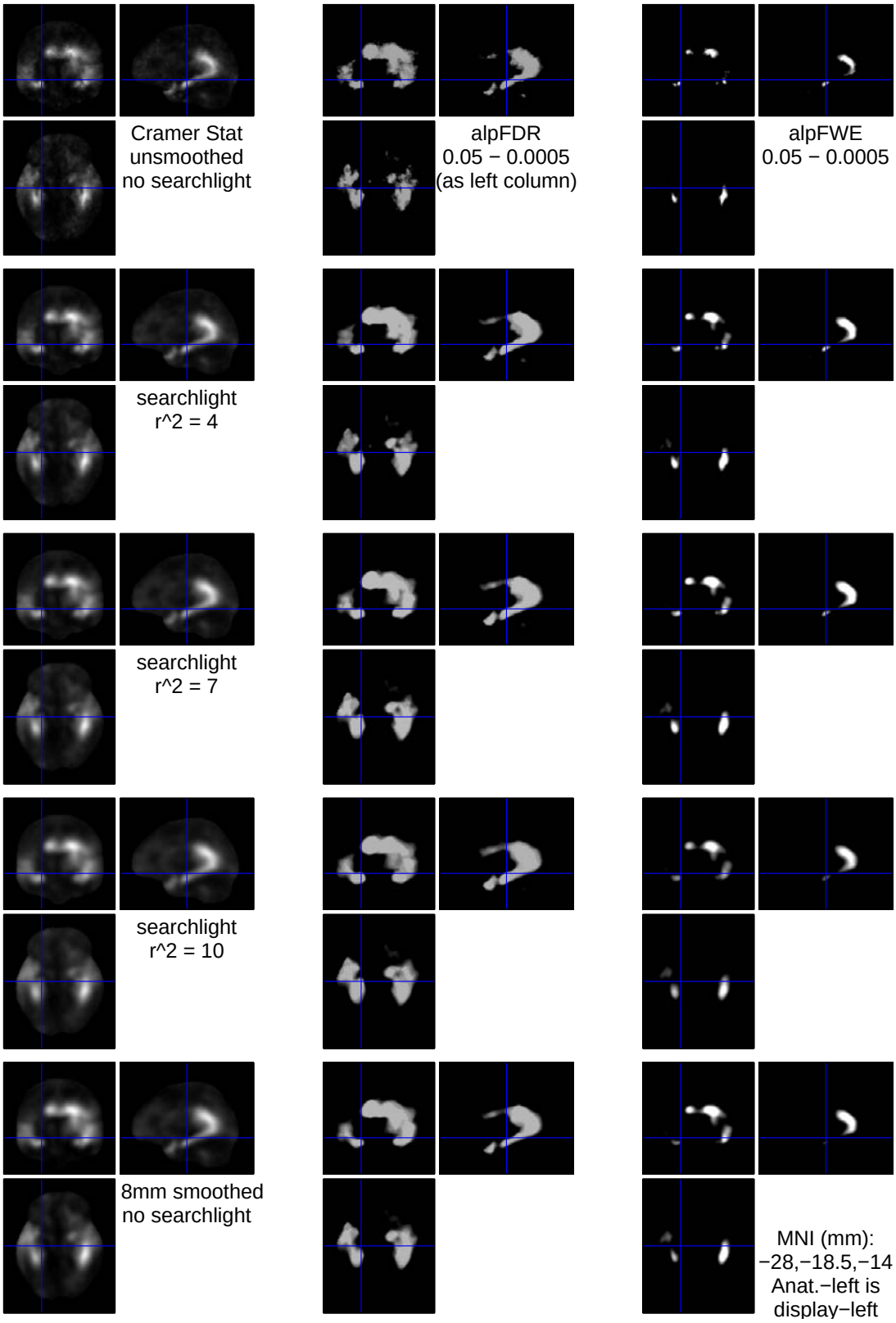
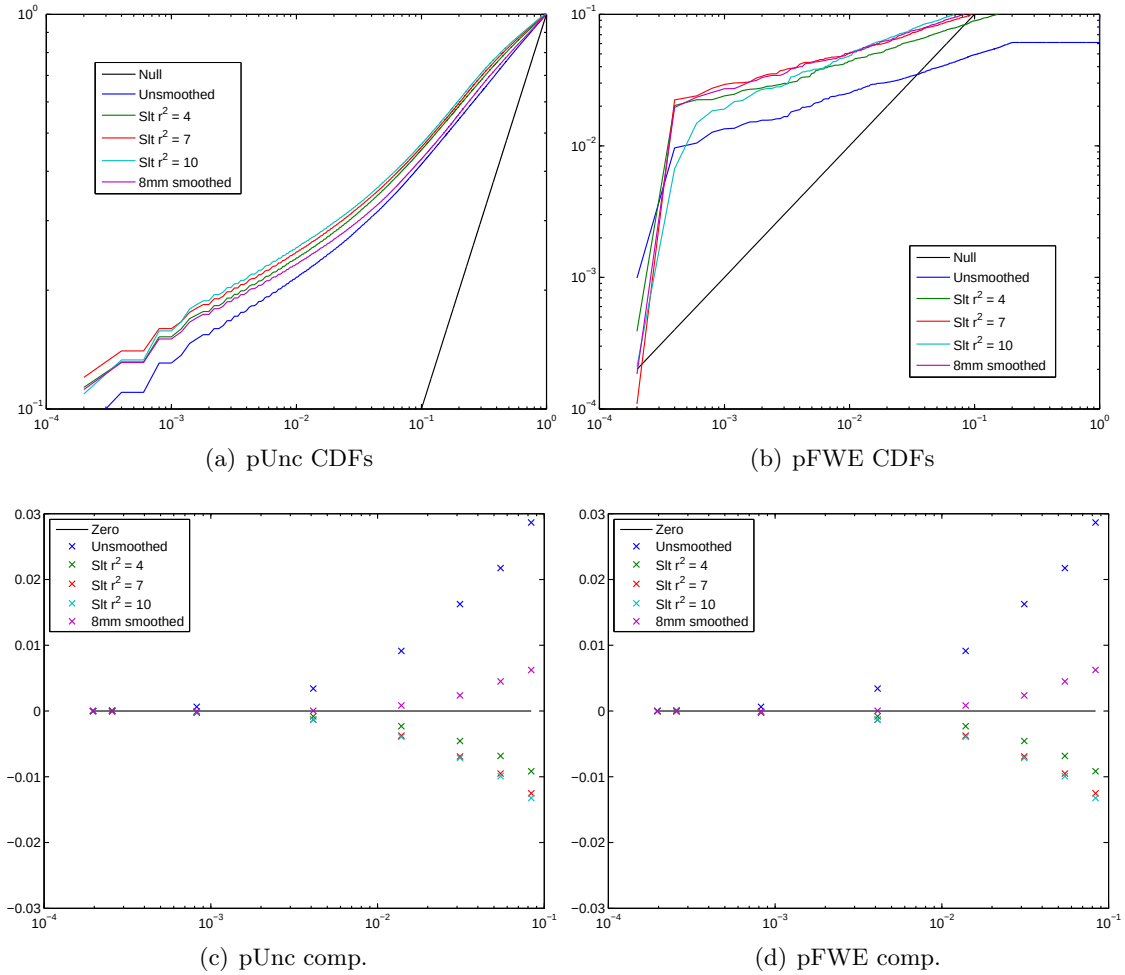


Figure 4.12: Statistical results for deformation-based morphometry using the Cramér test comparing the use of searchlight (middle three rows) to raw data (top row) or conventional smoothing (final row). P-values are displayed as absolute log p-values (brighter is more significant).



(a) pUnc CDFs (b) pFWE CDFs

(c) pUnc comp. (d) pFWE comp.

Figure 4.13: Comparison of searchlight and smoothing for DBM, in terms of (above) p-value CDFs, and (below) matched voxel p-value comparisons; for (left) uncorrected and (right) FWE-corrected p-values.

12-month interval data used above. Figure 4.14 overlays the Cramér statistic values and resultant log-transformed FDR p-values for the smallest (33 voxel) searchlight kernel considered on top of the corresponding results for the largest (144 voxel) kernel. As with the DBM results above, the larger searchlight kernel extends the regions of significance slightly, but loses some of the finer detail — much as conventional smoothing, but to a lesser degree.

Figure 4.15 presents maximum intensity projections of the FWE-corrected significant voxels for all six searchlight kernels considered. The MIPs are essentially the same, differing only slightly in the extent of their findings, while no distinct regions are present for some of the kernels but not others.

Figure 4.16 compares the observed powers of the different searchlight kernels, to each other and to the conventionally smoothed log-determinant. Unexpectedly, the conclusions differ depending on whether one focusses on uncorrected or FWE-corrected p-values. Uncorrected results show very similar power for 8 mm smoothing and for the 57-voxel ($r^2 = 5$) searchlight. However, when controlling FWE, it is the largest kernels that are closest to

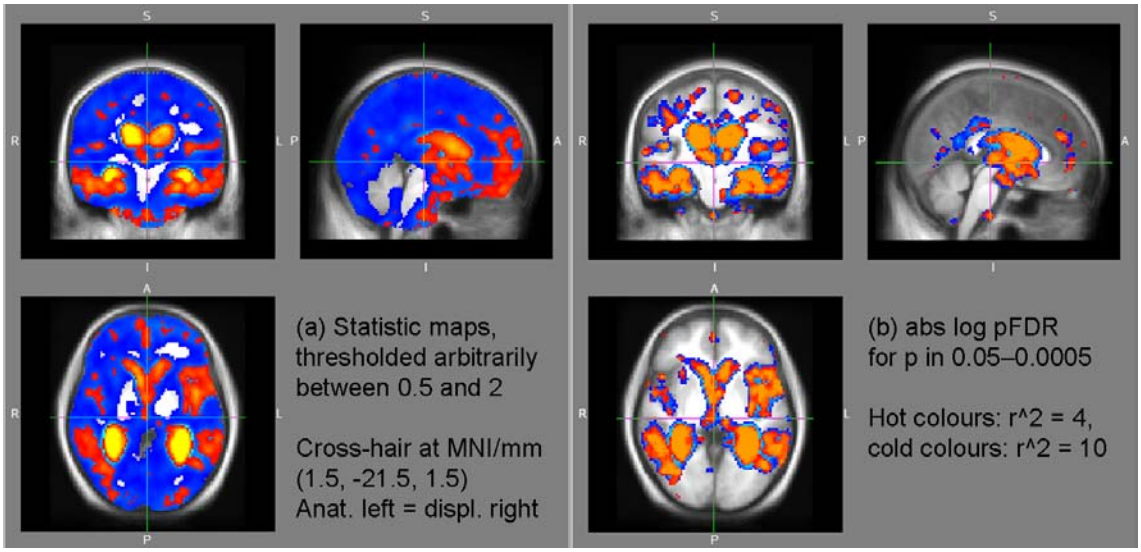


Figure 4.14: Results from searchlight Tensor-Based Morphometry on the log-transformed Jacobian determinant. Cramér test statistic and FDR-corrected p-values (shown on a logarithmic scale).

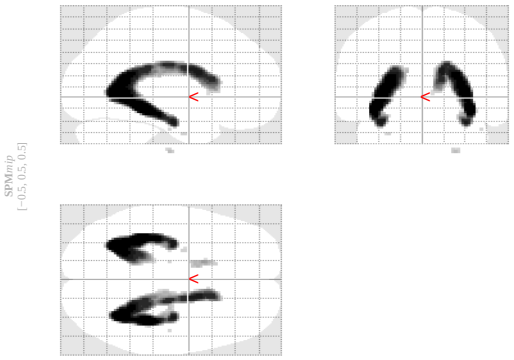
the power of smoothing (in fact, smoothing is the most powerful option for significance levels above about 0.003). This is of great importance, since it appears to be the smaller kernels ($r^2 \leq 6$) that preserve a similar amount of anatomical detail to smoothing; the very large kernels introduce excessive blurring.

As a very simple (in fact, greatly over-simplified) quantitative summary of panels (a) and (b) of figure 4.16, table 4.8 shows the numbers of voxels that survive the arbitrary $p < 0.05$ threshold under the three different levels of multiple comparison correction. The equivalent numbers for conventional Gaussian smoothing of the log-determinant lie between those of $r^2 = 4$ and $r^2 = 5$ in terms of uncorrected or FDR-corrected p-values, but are greater than any of the considered r^2 values when judged on FWE-corrected p-values.

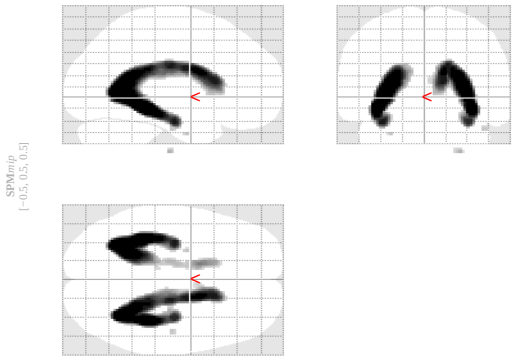
Searchlight r^2	Uncorrected	FDR corrected	FWE corrected
4	78099	48049	3366
5	87343	59514	4728
6/7	92725	66321	5566
8	95387	69924	6225
9	101320	77363	7041
10	104683*	81458*	7827
8 mm smoothing	80536	55586	9868*

Table 4.8: Numbers of supra-threshold voxels at $p < 0.05$ for the three different levels of correction, using different searchlight kernels on the log-determinant. Results with conventional smoothing are given for comparison. The maximum within each column is starred.

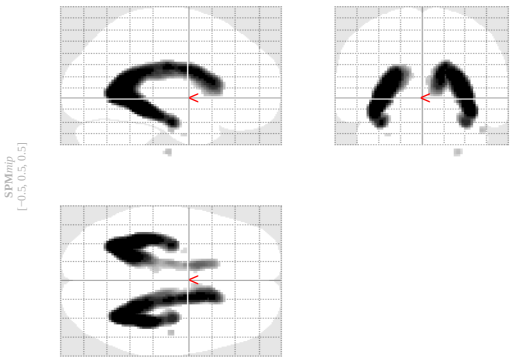
One potential explanation for the disappointing performance of the searchlight after FWE-correction (put forward by Thomas Nichols) is that searchlight can be seen as a form of adaptive smoothing, essentially taking advantage of the information in the voxels



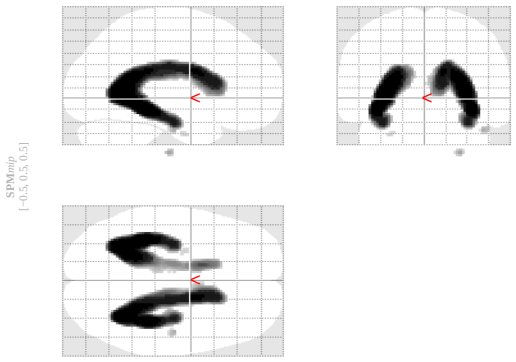
(a) squared radius = 4



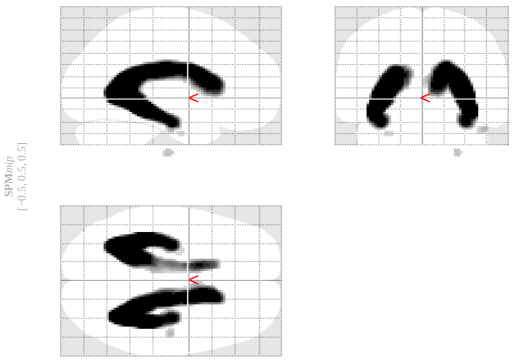
(b) squared radius = 5



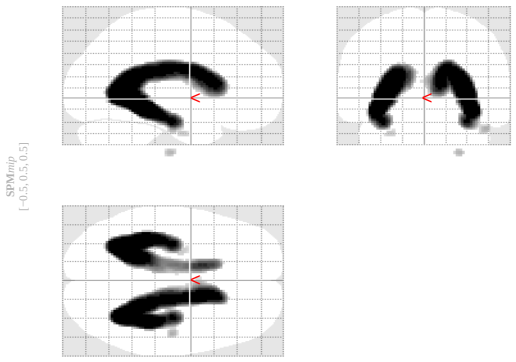
(c) squared radius = 6



(d) squared radius = 8

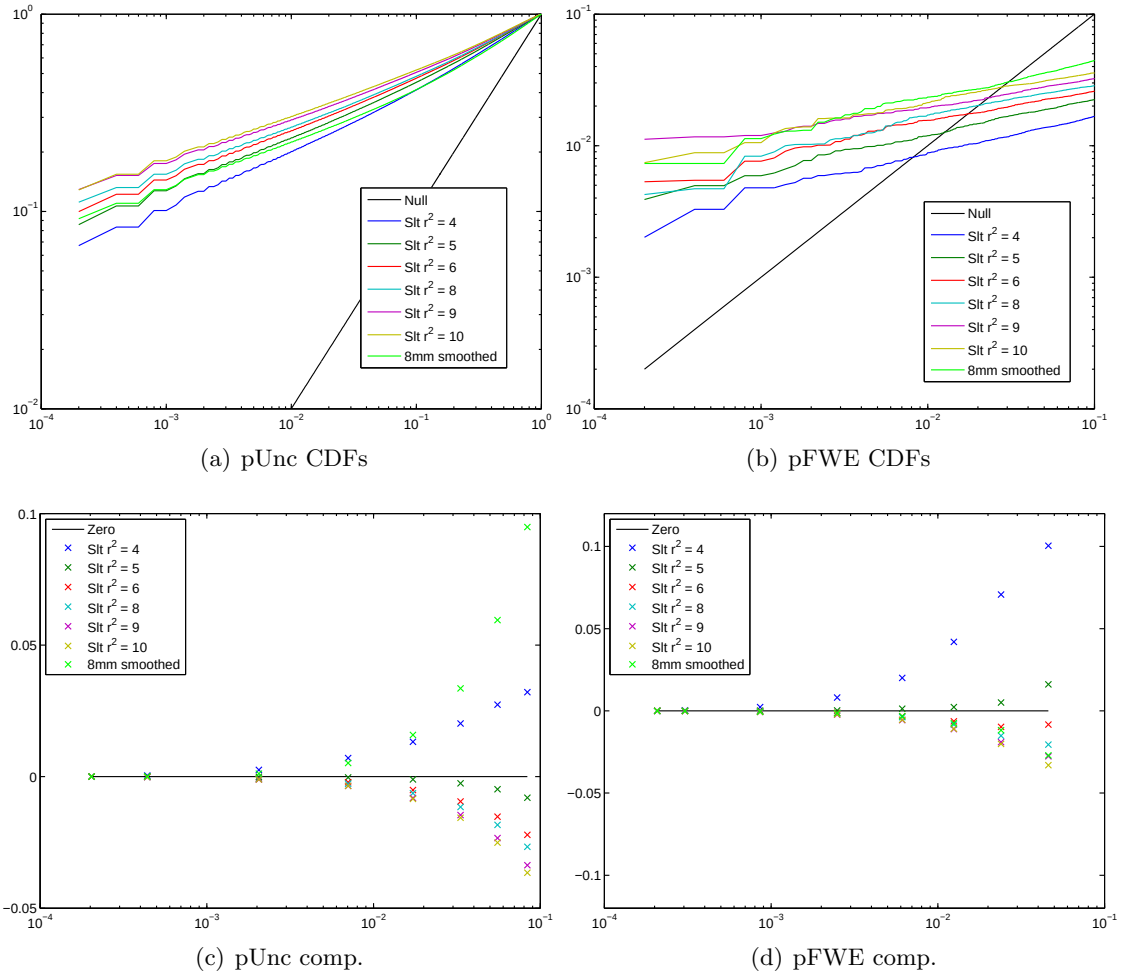


(e) squared radius = 9



(f) squared radius = 10

Figure 4.15: Maximum intensity projections of significant findings for searchlight TBM using the Cramér statistic. Absolute log p-values over the range $0.00005 < pFWE < 0.05$ are shown. Anatomical-left corresponds to display-left.



(a) pUnc CDFs (b) pFWE CDFs

(c) pUnc comp. (d) pFWE comp.

Figure 4.16: Comparison of searchlight and smoothing for TBM, in terms of (above) p-value CDFs, and (below) matched voxel p-value comparisons; for (left) uncorrected and (right) FWE-corrected p-values.

Searchlight r^2	FWHM_x	FWHM_y	FWHM_z	$\sqrt[3]{\prod_i \text{FWHM}_i}$
4	9.384	13.05	10.48	10.87
5	10.59	14.90	11.82	12.31
6/7	11.27	16.14	12.66	13.20
8	11.63	16.68	13.08	13.64
9	12.32	17.81	13.92	14.51
10	12.84	18.35	14.41	15.03
s.log det J	10.40	15.63	11.59	12.35

Table 4.9: Smoothness of searchlight TBM statistic images in terms of Gaussian Full-Width at Half-Maximum, in mm, as in figure 4.7.

included within the searchlight in a more complex way than simple Gaussian weighting. This adaptivity and complexity however, might be expected to increase the variability of the estimates at each voxel, leading to more outlying voxels and hence heavier tails in the permutation distribution of the maximum. This is somewhat speculative however, and further experimental work, probably including Monte Carlo simulation, will be necessary to understand this phenomenon. We report the estimated smoothness values for the statistic image in table 4.9 in an attempt to shed some light on the above difference. However, conventional smoothing results in a similar estimated smoothness to a searchlight kernel of $r^2 = 5$, which therefore fails to explain the relatively poorer FWE performance of the $r^2 = 5$ searchlight compared to 8 mm smoothing. On the basis of these (admittedly limited) results, the searchlight technique seems to offer no advantage to TBM when judged on FWE-corrected performance.

Searchlight over multiple scales

A potential problem with the searchlight is that even moderately large separation between the current voxel and its furthest neighbours requires an unreasonably large total number of voxels within the kernel. For example, the largest kernel here, with 147 voxels still only reaches a maximum radius of 3 voxels away from the centre. To reach 4, 5 and 6 voxel distances respectively requires 257, 515 and 925 voxels. While even the smallest kernel considered here ($r^2 = 4$) has almost as many voxels as the number of subjects (33 cf. 56). In an attempt to increase the spatial range of the searchlight analysis without unduly increasing the number of voxels being considered, we proposed to downsample the images. In particular, we suggest the use of Unser et al.'s spline-pyramids [42, 43], which provide optimal L_2 approximation, and allow easy upsampling of data.

Figure 4.17 uses a constant $r^2 = 4$ searchlight kernel, but with three different image resolution levels, 2 mm, 4 mm and 8 mm isotropic. Results are compared to conventional smoothing at the finest level. Disappointingly, this scale-space approach shows no evidence of benefit, for this particular data-set. No new locations become significant; originally significant regions extend to cover larger areas, but without respecting anatomical boundaries. We do not pursue this approach further here, since CDFs etc. seem uninteresting given the visually poor FDR and FWE results, however, we do not claim to have proven that the method has no potential; it must be applied to several other data-sets to fully characterise its merit.

4.4.4 Tensor-based morphometry

Results for five different tensor-derived measures over the 12-month interval are shown in figure 4.18 in terms of the unthresholded Cramér test statistic, and the thresholded FDR and FWE corrected p-values on a logarithmic scale. The most striking aspect of the statistic images is the general tendency for the higher dimensional measures to exhibit more widespread evidence of a group-difference between patients and controls. This is also supported by the FDR p-values. In particular, we observe that while the maximum eigenvalue extends some regions of significance beyond those present for the log-determinant, it also

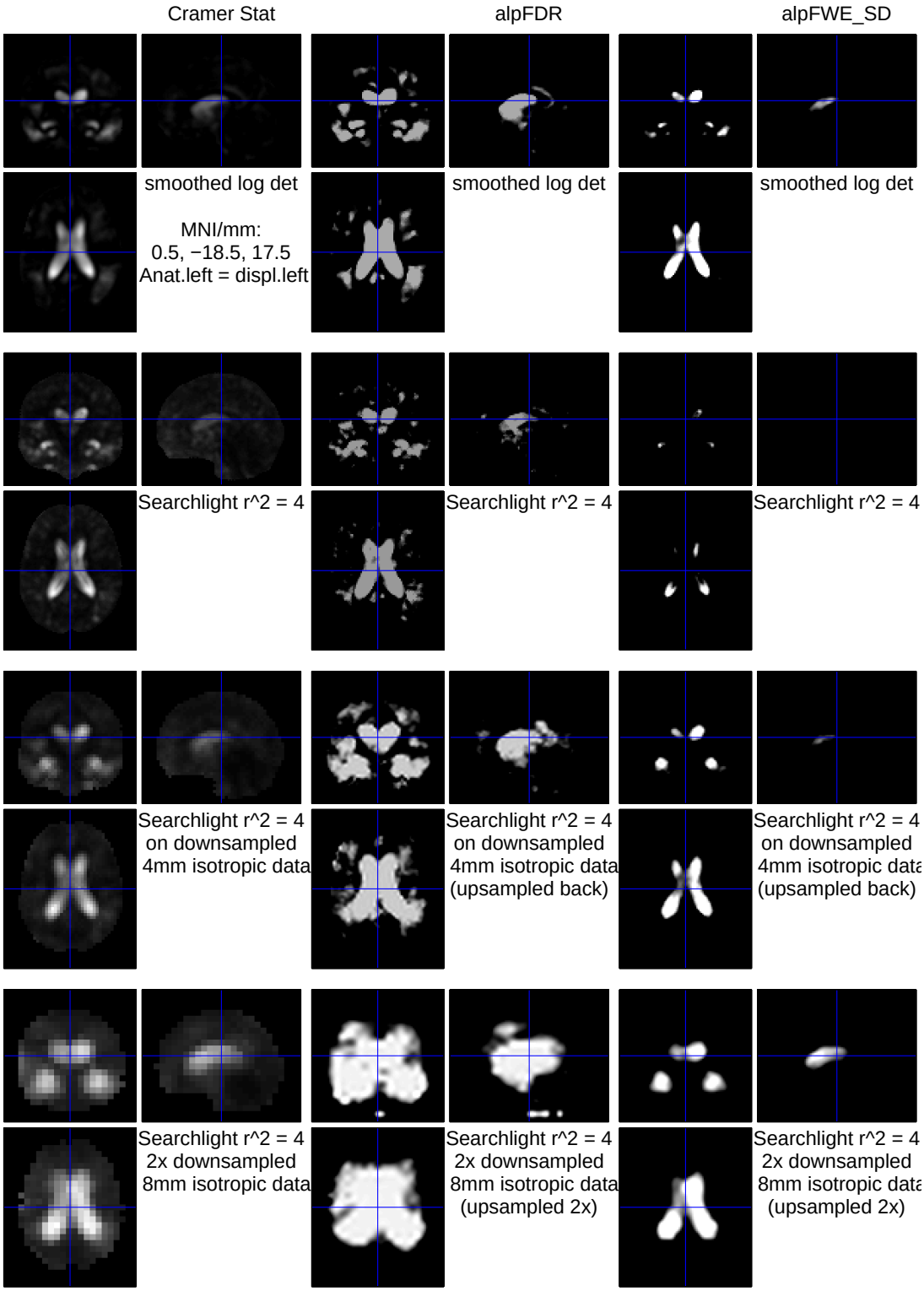


Figure 4.17: Statistical results for TBM using the Cramér test. Top row: conventional smoothing (8 mm FWHM); following rows, $r^2 = 4$ voxelsearchlight kernel, on original 2mm data, and two levels of spline-pyramid downsampled data. Statistic maps are shown at the downsampled resolution, but the FDR and FWE (log) p-values have been returned to the 2mm resolution through spline pyramid upsampling. P-values are displayed as absolute log p-values (brighter is more significant).

loses others. Multivariate testing of the set of eigenvalues appears to restore all these areas of lost significance, while adding additional voxels in clinically plausible locations including parts of the frontal cortex. Note that the the log-determinant is very similar (differing only in terms of smoothing) to the trace of the Hencky tensor (results not shown), which is the sum of the eigenvalues of H. This sum in turn is likely to be dominated by the largest of its components, the maximum eigenvalue, in voxels exhibiting anisotropic strain, which explains the overall similarity of the maps. Clearly, the (non-linear) transformations from J to $H = \operatorname{logm}\left(\left(J^T J\right)^{1/2}\right)$ and then to the eigenvalues can only reduce the total amount of information available, at the same time as the dimensionality, so it is unsurprising that the higher dimensional statistics show evidence of change over a greater number of voxels. These results, in terms of the statistic and FDR-corrected p-values (and also the unshown uncorrected p-values, from which the FDR-adjusted ones are monotonically derived) are in agreement with the results from the literature on generalised tensor-based morphometry [22, 23].

Interestingly, and counter-intuitively, it appears that the above trend has not been reproduced in the FWE-corrected results (presented here for the first time using generalised TBM). The third column of figure 4.18 shows very similar, anatomically reasonable, patterns of significant difference, but without any clear preference for the higher dimensional measures. In fact, the maximum intensity projections for FWE significance of the different results presented in figure 4.19, show clearly that the set of eigenvalues of H produces more widespread findings than either the complete set of unique elements of H or the full Jacobian tensor. (We note in passing, that the MIPs add the sixth and final measure from table 4.4, $\operatorname{tr}(J)$, to the five in figure 4.18, but that there is barely any visually discernible difference between this 'volume dilatation' and the volume change that it approximates, encoded in the log-determinant.)

To provide a simple quantitative summary of the above remarks, table 4.10 shows the numbers of voxels that survive the arbitrary $p < 0.05$ threshold using each p-value correction method.

Measure	Uncorrected	FDR corrected	FWE corrected
s.log det J	80536	55586	9868
smth trace J	78288	52981	9366
max e-val s.H	72958	45137	10936
e-vals s.H	134333	118735	20731*
smth H	136303	123816	13788
smth J	145945*	134802*	14526

Table 4.10: Numbers of supra-threshold voxels at $p < 0.05$ for the three different levels of correction, using various TBM measures. The maximum within each column is starred.

Avoiding the arbitrariness of any particular significance level, the p-value cumulative distribution functions, both without correction and with FWE correction, are shown in figure 4.20. The conclusions are essentially the same: the uncorrected results favour the full Jacobian followed by the Hencky tensor and its eigenvalues over the whole range of significance levels (from 0.0002, the reciprocal of the 5000 permutations performed, upward).

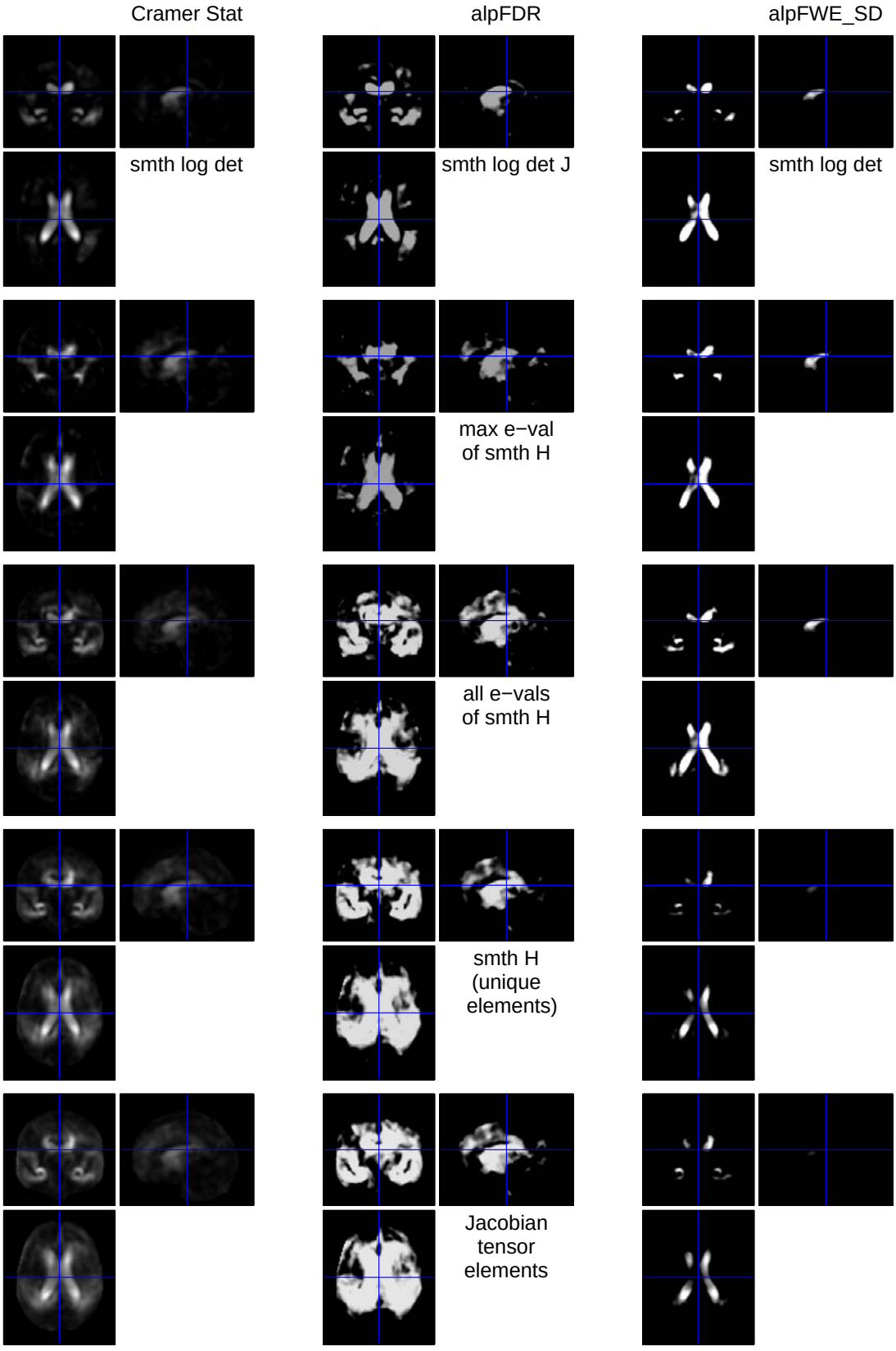
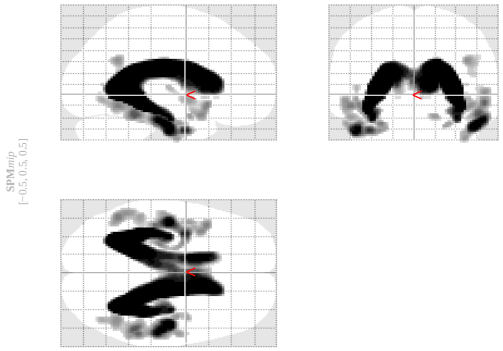
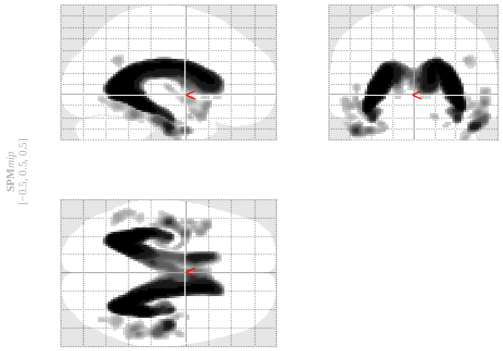


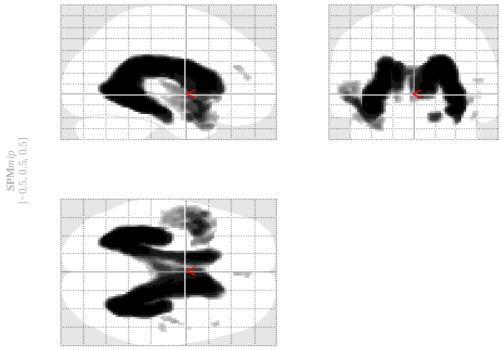
Figure 4.18: Statistical results for tensor-based morphometry, using the Cramér test, with 8 mm FWHM smoothing, on the measures from table 4.4, except $\mathrm{tr}\,(J)$, which is similar to $\det\,(J)$. P-values are displayed in the range 0.05–0.0005 as absolute log p-values (brighter is more significant). Anatomical-left is display-left. The cross-hairs are located at (0.5, -18.5, 17.5) mm MNI.



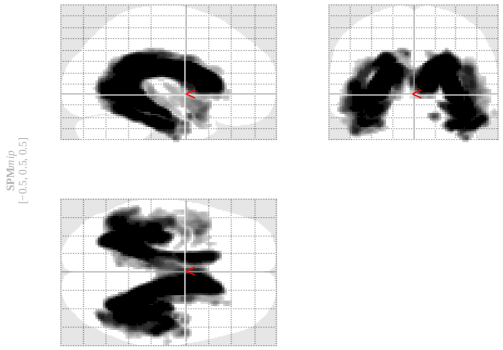
(a) Smoothed log det J



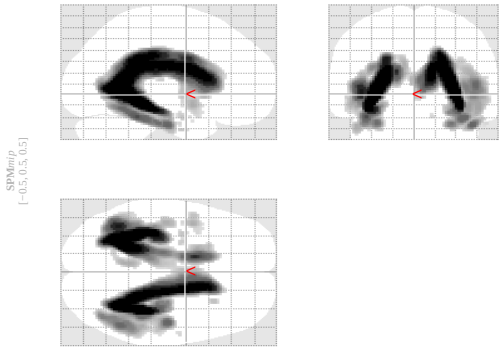
(b) Smoothed trace J



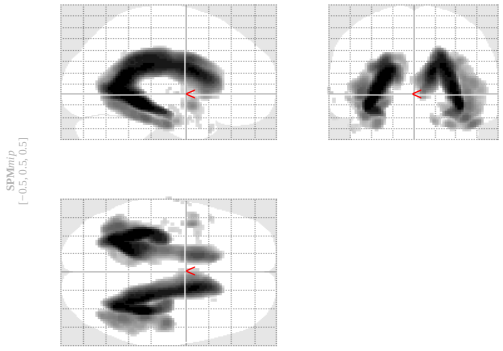
(c) Max e-val smoothed H



(d) All e-vals smoothed H



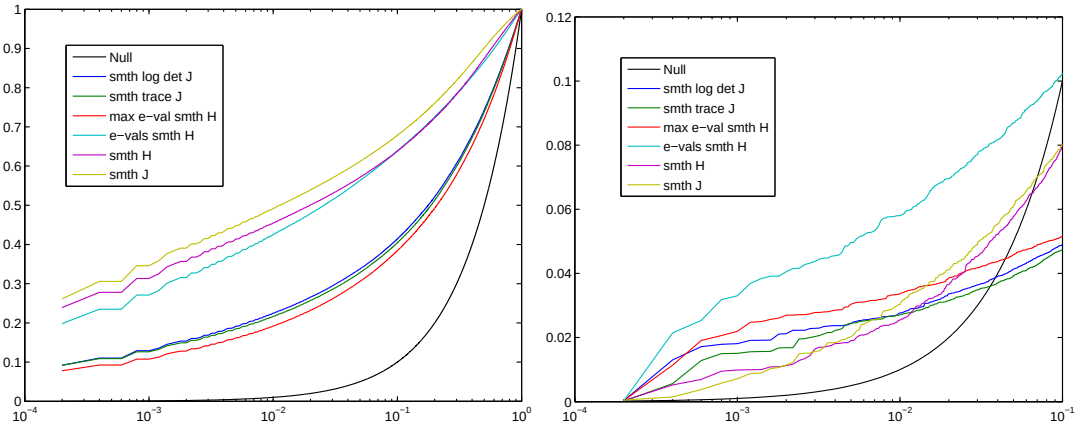
(e) Unique components of smoothed H



(f) Components of smoothed J

Figure 4.19: Maximum intensity projections of significant findings for TBM on the measures from table 4.4. Absolute log p-values over the range $0.00005 < pFWE < 0.05$ are shown. Anatomical-left corresponds to display-left.

The three univariate and three multivariate measures actually fall into two distinctly separate groups, with similar performance within each group. The FWE CDF curves are considerably more complex, with the relative order of the methods changing quite substantially between e.g. 0.005 and 0.05. Nevertheless, the set of eigenvalues is consistently the most powerful measure, while H and J perform quite badly at the most stringent p-values, below about 0.01, before crossing the univariate measures to become second only to the eigenvalues at about 0.05 (consistent with the numbers of supra-threshold voxels in table 4.10). Looking again at figures 4.18 and 4.19, the explanation seems to be that the higher dimensional measures have a broader but weaker pattern of significant voxels; giving them greater numbers at lenient thresholds, but becoming disproportionately less powerful at stricter levels of significance compared to the lower dimensional measures.



(a) pUnc CDFs (b) pFWE-SD CDFs

Figure 4.20: Statistical power of the different TBM measures illustrated via cumulative distribution functions of (a) uncorrected, and (b) FWE (step-down) corrected p-values.

False-discovery-rate p-value CDFs might also be of interest. For a particular method, the FDR p-values are monotonically related to the uncorrected ones, implying that the corresponding CDF curves will also be monotonically related. However, it is possible for different methods to exhibit a different relative ordering (e.g. a different method having the highest CDF curve at a particular level) in their uncorrected and FDR corrected CDFs. On sets of random p-values, this phenomenon can be quite dramatic, with the majority of levels showing different orders of 'methods', however, on the imaging results presented here, the FDR CDFs have shown nearly identical orderings of the methods to the uncorrected CDFs, and hence are omitted in the interest of space.

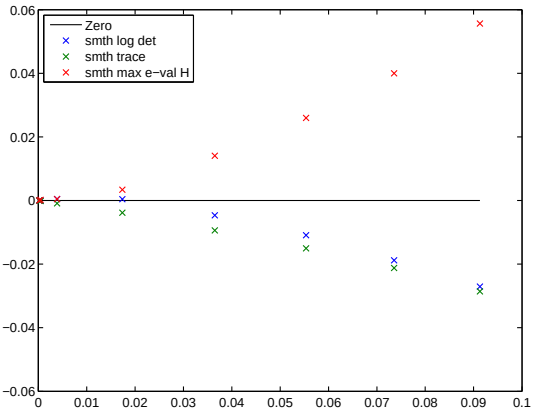
Moving from overall power to a comparison of sensitivity at corresponding voxels, figure 4.21 shows three separate plots of p-value differences with respect to (each plot's) mean, sorted by this mean value. This figure is based on uncorrected p-values, and shows patterns largely consistent with the results presented thus far. Focussing briefly on the choice between log-determinant and trace of the Jacobian, the means in panel (a) favour the trace, while the medians in (b) slightly favour the determinant. The CDFs in figure 4.20 showed no meaningful difference for uncorrected p-values but uniformly favoured the log-determinant to the trace in terms of FWE results. Panels (a-d) of figure 4.21

also introduce a new measure, the smoothed maximum eigenvalue of the Hencky tensor, instead of the maximum eigenvalue of smoothed H. This is included so that panels (a) and (b) compare three measures which are all smoothed at the end of their processing, however, the smoothed max eigenvalue performs very badly both in this comparison, and in panels (c) and (d), which provide a comparison of three eigenvalue-based measures, and is hence not considered further here (though other similar investigations related to the interaction of smoothing and preprocessing are pursued in section 4.4.4). The eigenvalue comparison shows the superiority of multivariate analysis of the set compared to both univariate analyses. The final two panels, (e) and (f), consider the multivariate measures, finding that higher dimensionality is associated with higher power.

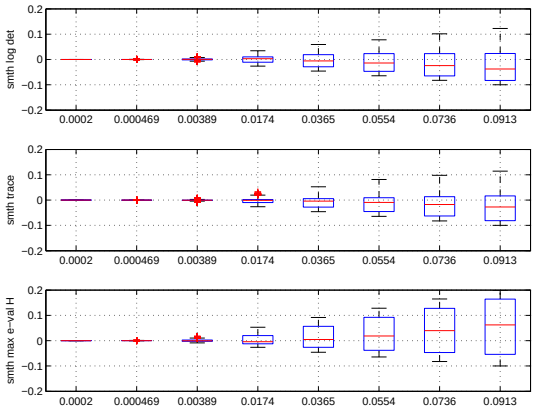
In figure 4.21, we have focussed on uncorrected p-values for method comparison, because, as discussed earlier (section 4.3.3) uncorrected p-values are individually valid and hence suited to this form of voxel-matched comparison. However, given the earlier conflicting conclusions regarding the best performing TBM measures, it seems prudent to briefly explore matched p-value comparison on the FWE p-values, even though additional complexity is added to the interpretation. Figure 4.22 presents equivalent plots for both uncorrected and FWE corrected p-values, for the most interesting TBM measures. Just as in the earlier figures, uncorrected results favour H and J to the univariate determinant, while FWE correction brings results for the tensors to near or below those of the scalar. The set of eigenvalues again appears to have some kind of optimal balance between the univariate and the multivariate, with at or near the best performance with or without correction. It will be important to investigate the reproducibility of this result in other data-sets; it is conceivable, for example, that the eigenvalues are optimal here with 56 subjects, but that the higher-dimensional strain tensor or Jacobian matrix could be preferable in larger studies.

Following the inconsistent findings from uncorrected and FWE-corrected results for the different measures, we now attempt to shed some light on two key aspects relating to FWE correction: the permutation distribution of the maximum, which underlies the corrected p-values; and the estimated smoothness of the statistical maps, which is related to the underlying family-wise error rate, since smoother fields of statistics imply fewer effective independent comparisons for which to correct. These two aspects are complementary in a sense, because smoothness relates to the spatial map as a whole, while the maximum distribution relates to the permutation-space, independent of the spatial location of voxels (for simplicity, we ignore the step-down procedure here).

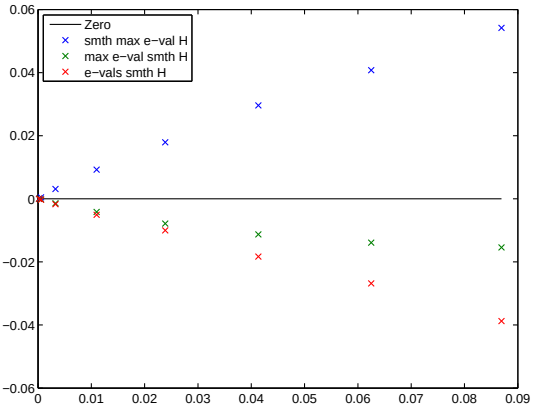
Considering first the permutation distribution of the image-wise maximal values from which the corrected p-values derive, figure 4.23 shows various illustrations of the distribution, and table 4.11 reports quantitative summary statistics. There is perhaps some evidence that the relatively poorer FWE performance of the higher dimensional measures might result from an increase in their maximum-distributions with respect to the maxima of their original identity-permutations, however, this hypothesis is far from being conclusively proven. For example, figure 4.23(c) provides the strongest support for this idea, with the full Jacobian followed by the Hencky tensor having the highest permutation maxima



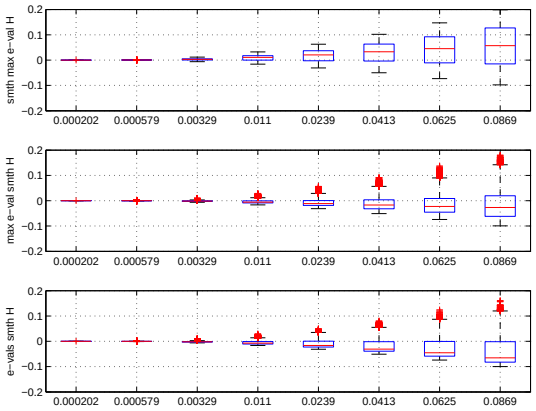
(a) Low dimensional TBM measures



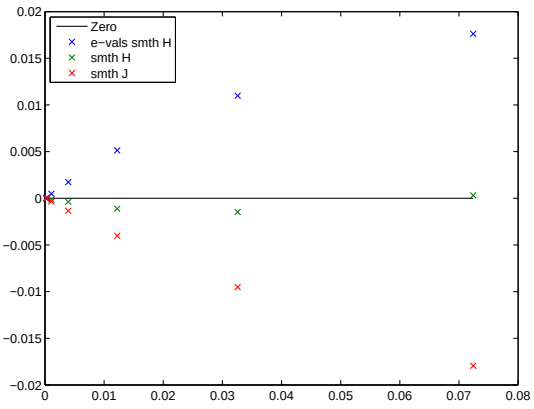
(b) Low dimensional TBM measures



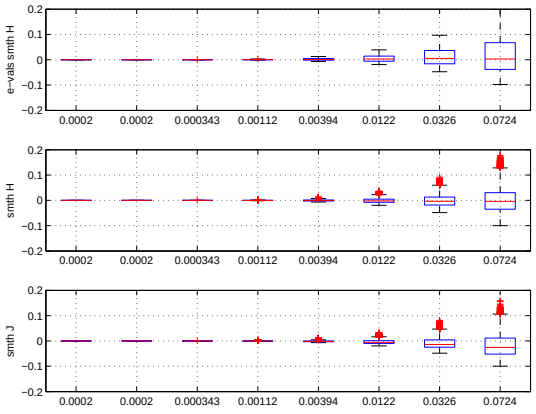
(c) Hencky eigenvalue TBM measures



(d) Hencky eigenvalue TBM measures

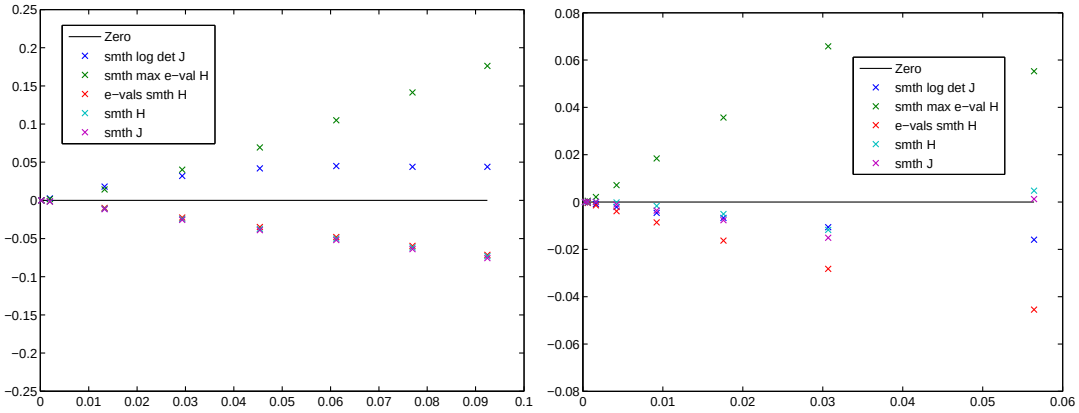


(e) High dimensional TBM measures



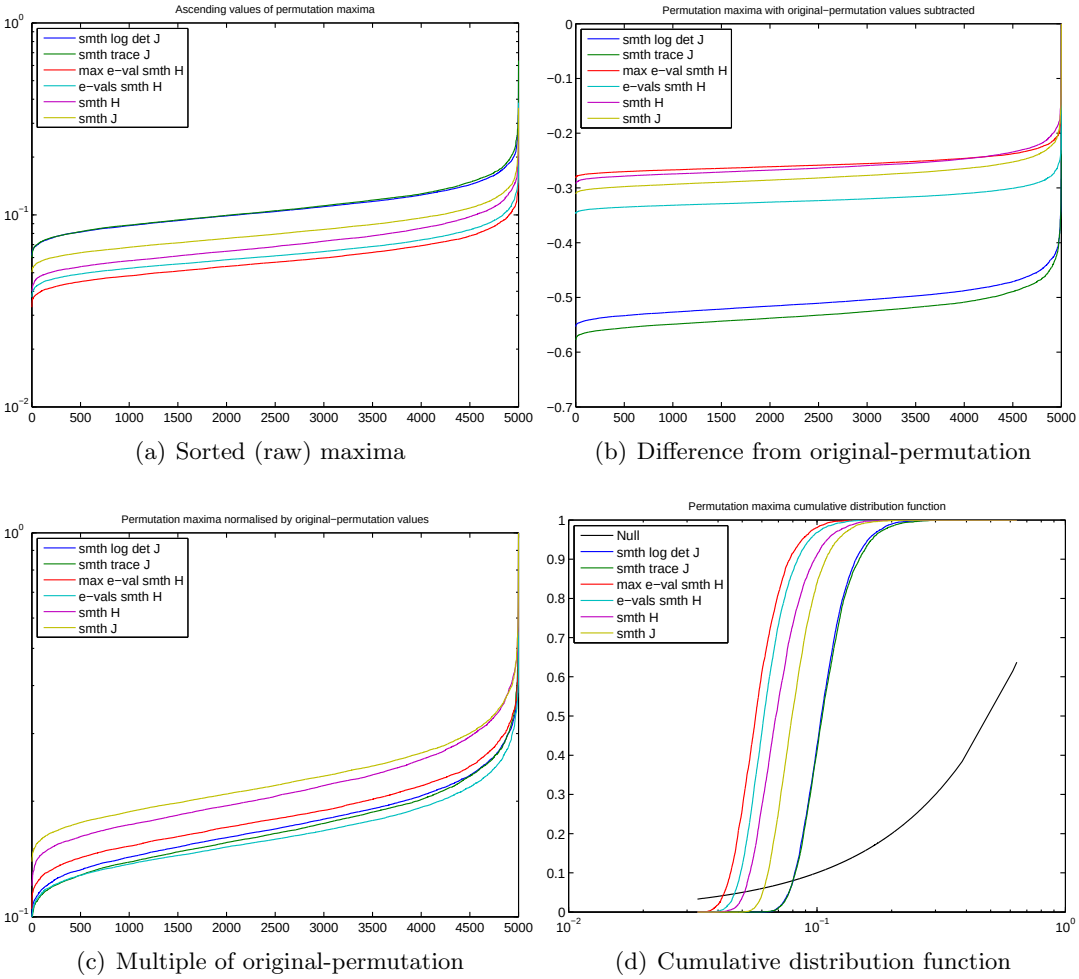
(f) High dimensional TBM measures

Figure 4.21: Statistical power of the different TBM measures from table 4.4 illustrated via voxel-matched comparison of their uncorrected (permutation-based) p-values. The p-values are compared (to their group means) in separate groups of (a-b) 'low' and (e-f) 'high' dimensionality, with a group of eigenvalue based measures (c-d) used to bridge the gap from low to high dimensional. The left column shows means, while the right shows boxplots (with medians in their centres, as usual). Lower points correspond to lower (more significant) p-values.



(a) Uncorrected p-values (b) FWE-SD p-values

Figure 4.22: Statistical power of the some measures from fig. 4.21 illustrated via voxel-matched comparison of their uncorrected and FWE (step-down) corrected p-values.



(a) Sorted (raw) maxima (b) Difference from original-permutation

(c) Multiple of original-permutation (d) Cumulative distribution function

Figure 4.23: Illustrations of the 5000 values of the maximum statistic from each permutation, for the different TBM measures. (a) shows the maxima sorted into ascending order. The identity permutation produced the highest maximum values for each method, these six values have been used to standardise the permutation distribution by (b) subtraction and (c) division. The cumulative distribution function for the raw maxima is shown in (d) with a uniform CDF in black.

relative to their original maxima, and the Jacobian and trace being nearer the bottom. However, we cannot argue that this form of normalisation is more appropriate than that in figure 4.23(b), which shows a less easily interpretable ordering of the measures. Note that the CDF, figure 4.23(d) shows essentially the same pattern as the sorted raw distributions in fig. 4.23(a), with the top-to-bottom ordering of the measures in these two figures simply mirroring each other; if the CDF is produced with the normalisations in figures 4.23(b) or (c), it similarly reflects their orderings (graphs omitted for brevity).

In table 4.11, the most helpful summary statistic appears to be the kurtosis, in particular, the difference in kurtosis with and without the identity permutation is greatest for the set of eigenvalues, and larger for the univariate measures than for H or J, which is consistent with the relative FWE performance of the different methods illustrated in figure 4.20(b). It makes intuitive sense that this kurtosis difference is closely related to FWE performance, since it reflects the degree to which the original labelling's observed statistic is in the tails of the permutation distribution — exactly the basis of the FWE p-values. Unfortunately though, the table seems to offer little in the way of extra insight. One plausible a priori hypothesis was that the higher-dimensional measures would create greater potential for variability and/or for outlying extrema in the statistics. However, the standard deviation, skewness, and kurtosis all fail to show a consistent pattern with increasing dimensionality, implying that the phenomenon is more complicated than this simple hypothesis.

Measure	Mean	$\text{Mean}_{\setminus I}$	Stdev	$\text{Stdev}_{\setminus I}$	Skew	$\text{Skew}_{\setminus I}$	Kurt	$\text{Kurt}_{\setminus I}$
s.log det J	0.1097	0.1096	0.0282	0.0273	2.632	1.650	27.80	8.283
smth trace J	0.1113	0.1112	0.0303	0.0294	2.716	1.848	26.43	9.493
max e-val s.H	0.0596	0.0596	0.0147	0.0142	2.370	1.456	23.98	6.406
e-vals s.H	0.0645	0.0645	0.0158	0.0151	3.057	1.601	39.49	6.992
smth H	0.0729	0.0729	0.0197	0.0194	2.023	1.662	12.83	7.363
smth J	0.0837	0.0836	0.0204	0.0200	2.153	1.753	15.03	8.833

Table 4.11: Moment-based statistics (mean, standard deviation, skewness and kurtosis) summarising the permutation distribution of the maximum statistic, with and without the original identity-permutation, for the different TBM measures.

Measure	FWHM_x	FWHM_y	FWHM_z	$\sqrt[3]{\prod_i \text{FWHM}_i}$
s.log det J	10.40	15.63	11.59	12.35
smth trace J	10.46	15.82	11.76	12.49
max e-val s.H	11.60	17.30	13.17	13.83
e-vals s.H	12.72	17.97	13.70	14.63
smth H	12.94	20.03	14.73	15.63
smth J	13.45	20.31	15.15	16.06

Table 4.12: Smoothness of TBM statistic images in terms of Gaussian Full-Width at Half-Maximum, in mm, as in figure 4.7.

Approximate estimates of the smoothness of the data, in terms of the FWHM in mm of a Gaussian kernel, are presented in table 4.12. A grossly over-simplified approach has been

used to compute these values; they are estimated directly from the Cramér statistic images themselves, using the `3dFWHMx` program as used in section 4.4.3. A major improvement, left for future work, would be to estimate the smoothness of the noise, without the anatomical structure present in the statistic image itself, by using the method of Worsley [29] on the full set of (multivariate) residuals. Along all axes, the smoothness increases monotonically with increasing dimensionality of the TBM measures. This is consistent with subjective visual inspection of the statistic images in figure 4.18, but does not help to explain the poorer FWE performance of the higher-dimensional measures, since greater smoothness should indicate a less severe multiple comparison problem, and hence less need for the FWE correction to lower the uncorrected significance.

The failure of either smoothness or the maximum distribution to account for the discrepancy between uncorrected/FDR and FWE performance of the multivariate measures motivates further research. It will be important to see if this phenomenon is replicated using other data-sets in the future; for now, note that we already found the same behaviour using a different method on the same data, in terms of the size of searchlight kernels in section 4.4.3. We will now also investigate this effect on the six-month data, though reproducibility here will admittedly be much less compelling than on an entirely separate data-set.

TBM results for six-month data

Clinically, there is great interest in evaluating neurodegeneration over shorter intervals, with obvious motivations such as earlier diagnosis or more rapid detection of drug treatment effects. Methodologically, however, the length of baseline–follow-up interval should be immaterial, and therefore results are presented only briefly here. The chief technical question here is whether the six-month interval exhibits the same inconsistency between uncorrected and FWE-corrected results in terms of comparing the different TBM methods. Figure 4.24 presents the statistics, FDR, and FWE (step-down) corrected p-values, for a selection of tensor-based measures having dimensionalities 1, 3, 6 and 9. Reassuringly, the general pattern of findings is similar, which can be seen most clearly by comparison of the unthresholded statistic images with those from figure 4.18.

As expected, significance is lower for all results — maximum intensity projections given in figure 4.25 clearly show much smaller areas surviving $pFWE < 0.05$. Interestingly, the six-month data seem to exhibit the same pattern of higher-dimensional measures being favored by uncorrected significance but not FWE results, with the two highest dimensional measures (H and J) having the least corrected supra-threshold voxels of the four measures (counts are given in table 4.13). In fact, in terms of the areas surviving $pFWE < 0.05$, the phenomenon is even more pronounced over the shorter interval. It seems that the higher-dimensional data increase the number of voxels that appear prominent in the statistic images while slightly decreasing the highest statistic values, and similarly, produce more wide-spread but slightly reduced patterns of FWE significance. Over 12 months, enough of the wide-spread areas meet $pFWE < 0.05$ to make the higher dimensional measures appear comparable to or better than the scalar measures. With the smaller changes occur-

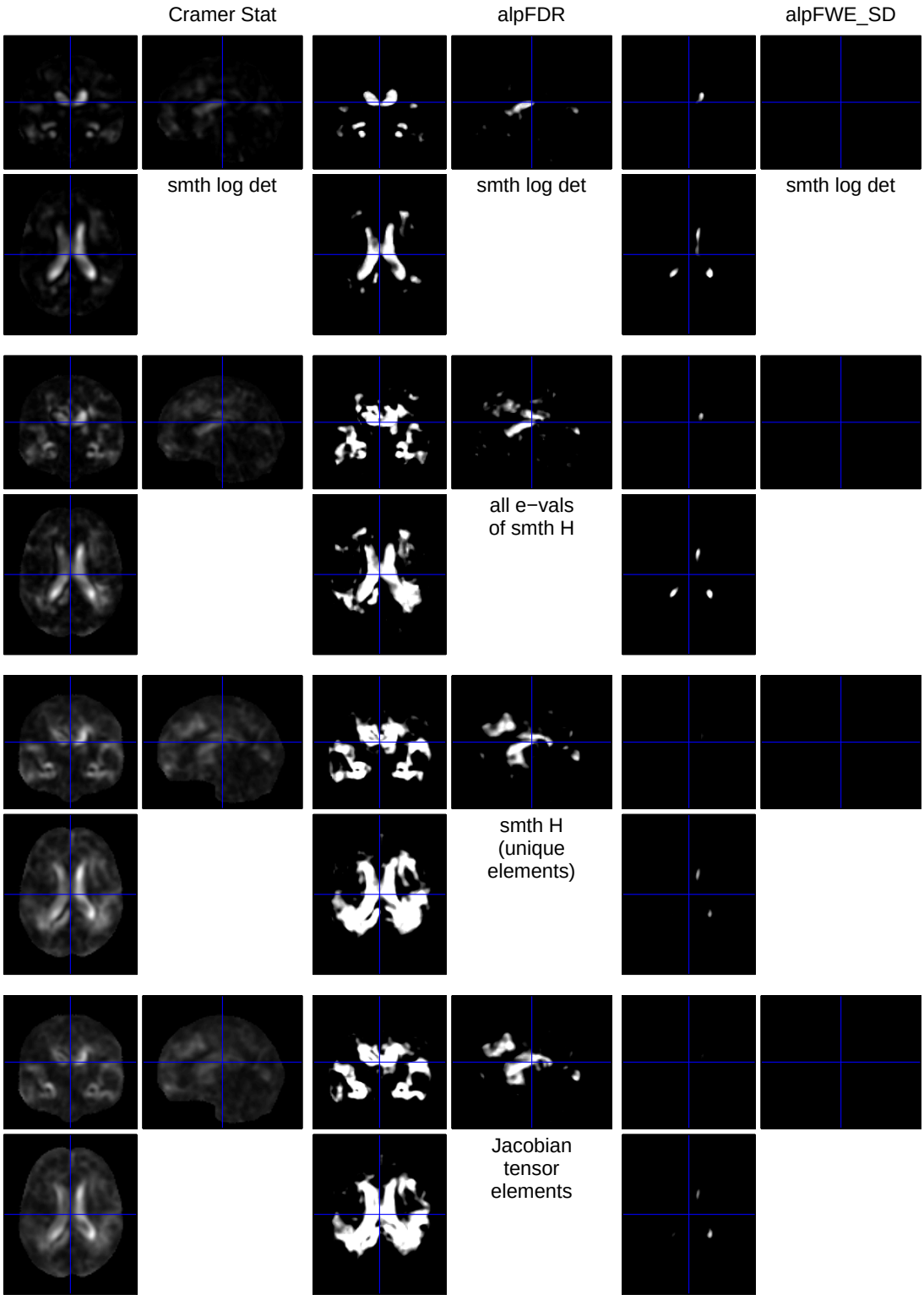
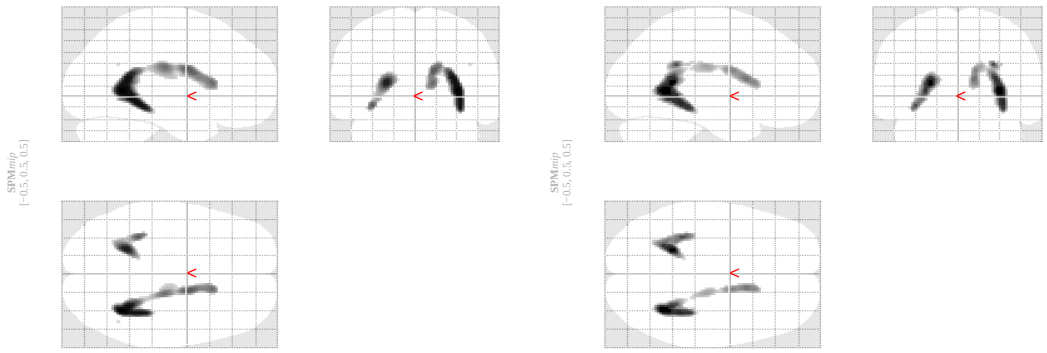


Figure 4.24: Statistical results for tensor-based morphometry over the six-month interval. Orientation and location of views matches figure 4.18, but here, the p-values are displayed using the more lenient range 0.1–0.01 (again as absolute log p-values).

ring over six months, the general reduction of significance with increasing dimensionality has a more noticeable detrimental impact on the number of voxels meeting the chosen threshold.

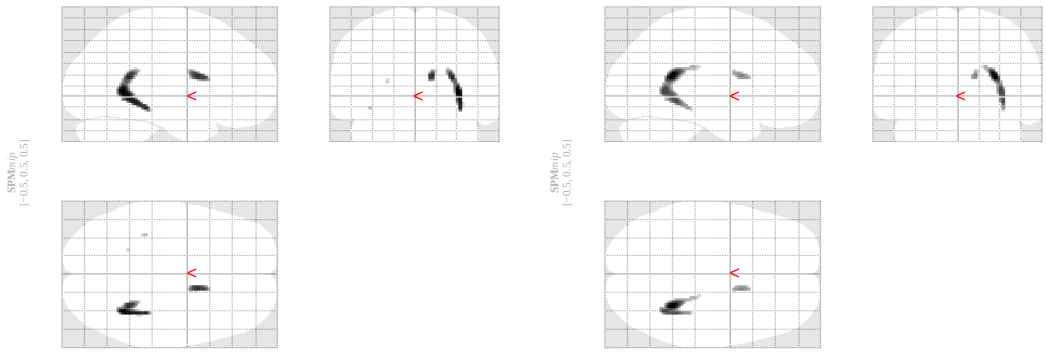
Measure	Uncorrected	FDR corrected	FWE corrected
s.log det J	43431	14514	1248*
e-vals s.H	70813	37410	1229
smth H	80098*	51769*	197
smth J	79491	50721	210

Table 4.13: Numbers of supra-threshold voxels at $p < 0.05$ for the three different levels of correction, using the TBM measures evaluated over the six-month interval. The maximum within each column is starred.



(a) Smoothed log det J

(b) All e-vals smoothed H



(c) Unique components of smoothed H

(d) Components of smoothed J

Figure 4.25: Maximum intensity projections of significant findings for TBM over the six-month interval. Details as for figure 4.19 (including the p-value range, which is consistent with the 12-month data, unlike figure 4.24 above, which relaxed the thresholds).

Regarding detailed comparisons based on p-values for the six-month interval, figures are not presented here, but the overall pattern of findings is similar to the longer time-period. In particular, uncorrected p-value CDFs favour the full Jacobian, followed by the Hencky tensor, the eigenvalues, and then the determinant, while FWE-corrected results favour the eigenvalues, followed very closely by the determinant, with H and J performing

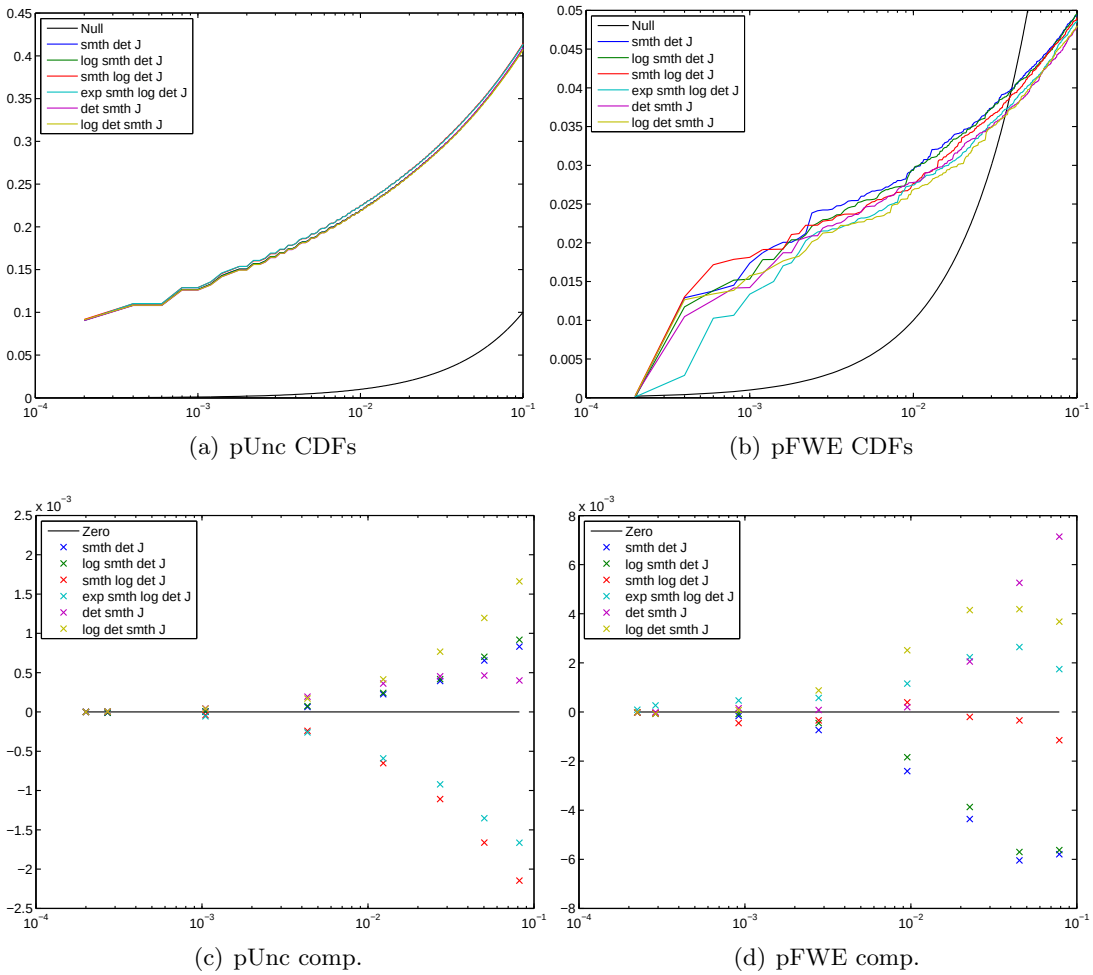
considerably worse, and similarly to each other.

The estimated smoothness values for the statistic images show a similar pattern to the corresponding results in table 4.12, but at lower levels of smoothness. The geometric mean FWHM for the four methods considered here are 10.11, 10.93, 12.07 and 12.04 mm. As with the 12-month data, the smoothness values are larger in the anterior-posterior direction (the separate x, y, and z geometric means over the four methods are 9.60, 13.31, and 11.16 mm). It seems unlikely (though possible) that the true smoothness of the underlying random field for the six-month data should be lower than for the longer interval, which adds weight to the suggestion above that this form of smoothness estimation directly from an image is unreliable, and should be replaced with residual-based estimation [29].

Methodological subtleties

Figure 4.26 and table 4.14 compare six different TBM options for the combination of smoothing and logarithmic transformation of the determinant of the Jacobian. Somewhat surprisingly, there is no visually discernible difference between these methods in terms of either the Cramér statistic maps or (absolute log) p-value maps, which are therefore not shown. Similarly, the CDFs for the uncorrected p-values are virtually identical. The voxel-matched comparison of uncorrected p-values given in figure 4.26(c) slightly favours the smoothed log determinant, followed by its exponentiated version. The simplest option, the smoothed determinant, is slightly worse than the average of the six measures. In terms of FWE corrected p-values, the CDF in figure 4.26(b) shows a changing pattern over the different thresholds, with the smoothed log determinant most powerful at significance levels below about 0.002, but with the simpler smoothed determinant out-performing it at more lenient thresholds. The exponentiated smoothed log determinant that was among the best methods in terms of uncorrected p-values is one of the two worst methods based on FWE performance. Voxel-matched comparison of FWE p-values shows the smoothed log determinant close to the average of the methods, while the untransformed smoothed determinant and the theoretically less-appealing log smoothed determinant appear to be superior. These results are somewhat counterintuitive, though note that the differences are very small (the y-scale on the p-value comparisons is considerably smaller than for other similar comparisons presented in this chapter. Given the theoretical appeal of the smoothed log determinant, and its popularity in the literature, there is insufficient evidence at this stage to suggest that it should be replaced by any of the other measures.

Equivalent smoothing options for the multivariate strain tensor and its eigenvalues are presented in figure 4.29 and table 4.15. It is immediately clear from both the uncorrected and FWE-corrected results that the eigenvalues are more powerful when derived from a smoothed tensor, than when smoothed directly themselves. Figure 4.28 presents maps of the statistic values and p-values, which reinforce this finding, and additionally suggest that the eigenvalues of the smoothed tensor result in a sharper, more anatomically refined set of results. Smoothness estimates of the Cramér statistic maps using `3dFWHMx` support this, with geometric FWHM values of 14.63 and 16.12 mm respectively. It is interesting in itself that the differences should be so dramatic regarding the order of the smoothing- and



(a) pUnc CDFs (b) pFWE CDFs

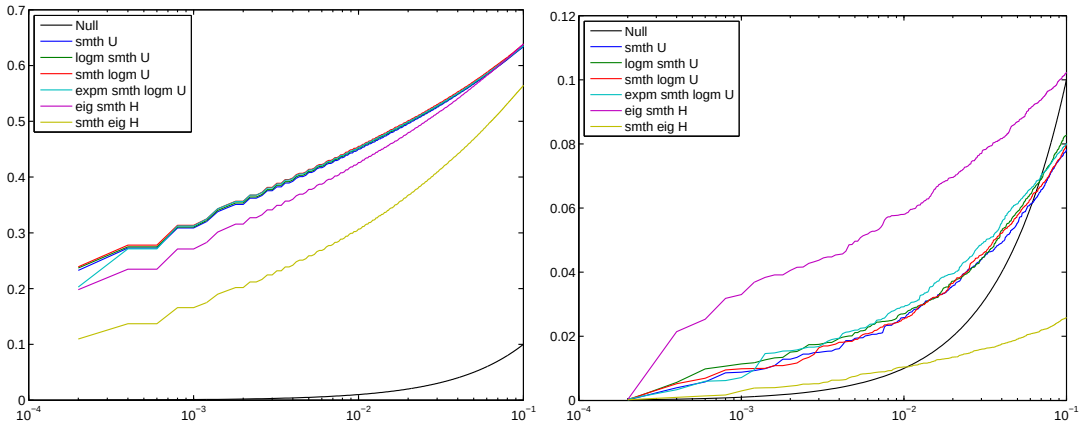
(c) pUnc comp. (d) pFWE comp.

Figure 4.26: Comparison of the smoothing/preprocessing options for the scalar determinant TBM measure, in terms of (above) p-value CDFs, and (below) matched voxel p-value comparisons; for (left) uncorrected and (right) FWE-corrected p-values.

Measure	Uncorrected	FDR corrected	FWE corrected
smth det J	79161	53931	10048*
log smth det J	79157	53888	9929
smth log det J	80536*	55586*	9868
exp smth log det J	80407	55497	9591
det smth J	79223	53757	9384
log det smth J	78874	53705	9489

Table 4.14: Numbers of supra-threshold voxels at $p < 0.05$ for the three different levels of correction, for the different options of preprocessing $\det(J)$. The maximum within each column is starred.

eigenvalue-operators, in contrast to the relatively minor differences found for interchanging the order of smoothing and scalar logarithm above (or even smoothing and matrix logarithm for the tensor measures shown later in figure 4.29). It is understandable that the determinant showed small differences, because the scalar logarithm operation is a simple monotonic one. However, the matrix logarithm might have been expected to make a greater difference.



(a) pUnc CDFs (b) pFWE CDFs

Figure 4.27: Comparison of the smoothing/preprocessing options for the multivariate TBM measures, in terms of p-value CDFs, for (left) uncorrected and (right) FWE-corrected p-values.

Measure	Uncorrected	FDR corrected	FWE corrected
smth U	135271	122589	13192
logm smth U	135606	123160	14058
smth logm U	136303*	123816*	13788
expm smth logm U	135754	123379	14628
eig smth H	134333	118735	20731*
smth eig H	111648	84696	4568

Table 4.15: Numbers of supra-threshold voxels at $p < 0.05$ for the three different levels of correction. For the multivariate strain tensor measures, including eigenvalues. The maximum within each column is starred.

Figure 4.29 focusses on the full strain tensor. With uncorrected p-values, the comparison favours the smoothed Hencky tensor, which has the greatest theoretical support; the simpler smoothed U (recall this is equivalent to testing the Biot strain tensor) performs the worst. For the FWE-corrected results, the Biot tensor remains poor, but the Hencky tensor becomes almost as bad, with the best measure now being the exponentiated smoothed log-transformed tensor. While further work is clearly required — not least to thoroughly search for any potentially misleading aspects of comparing FWE p-values instead of uncorrected ones — this finding is interesting in relation to results reported by Whitcher et al. [16]. With unsmoothed DTI data,[35] Whitcher et al. found that direct analysis of the

[35] As an aside, it would be interesting to investigate whether smoothing, perhaps in the log-Euclidean

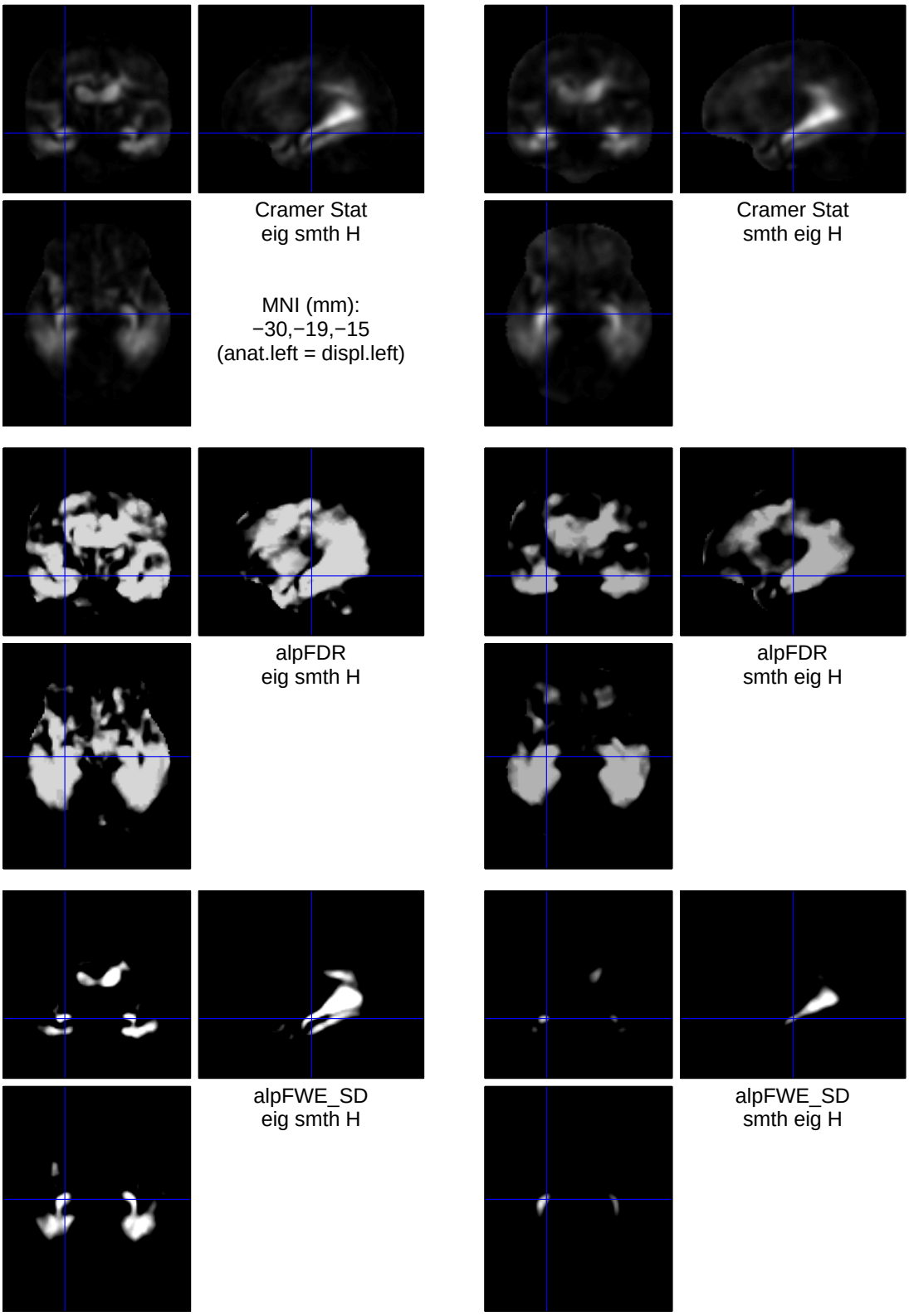
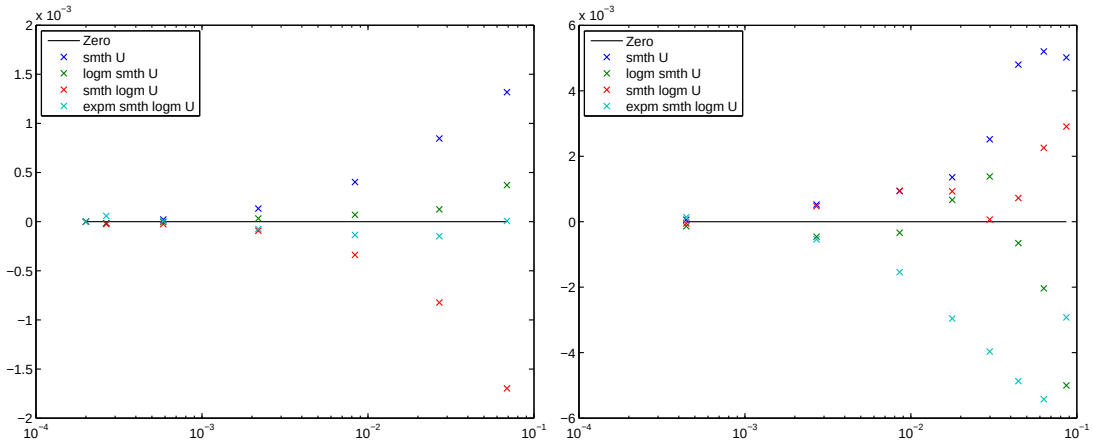


Figure 4.28: Statistical results for tensor-based morphometry, using the eigenvalues of the Hencky strain tensor, with two different smoothing options: left, eigenvalues of the smoothed H; right, smoothed eigenvalues of H. Top-to-bottom: Cramér statistic, FDR p-values, and FWE p-values. P-values are displayed in the range 0.05–0.0005 as absolute log p-values (brighter is more significant).

tensors preferable to log-Euclidean analysis; however, here, with the need for a smoothing step, it seems that log-Euclidean smoothing followed by matrix exponentiation improves upon results from direct analysis. However, finally for this particular study, we remark that the correction-processing interaction question is of the greatest importance, since uncorrected (and FDR-corrected) p-values in table 4.15 actually show the log-Euclidean analysis to be the most powerful approach, by a small margin.



(a) pUnc comp. (b) pFWE comp.

Figure 4.29: Comparison of the smoothing/preprocessing options for the multivariate strain tensor measures, including eigenvalues, in terms of matched voxel p-value comparisons; for (left) uncorrected and (right) FWE-corrected p-values.

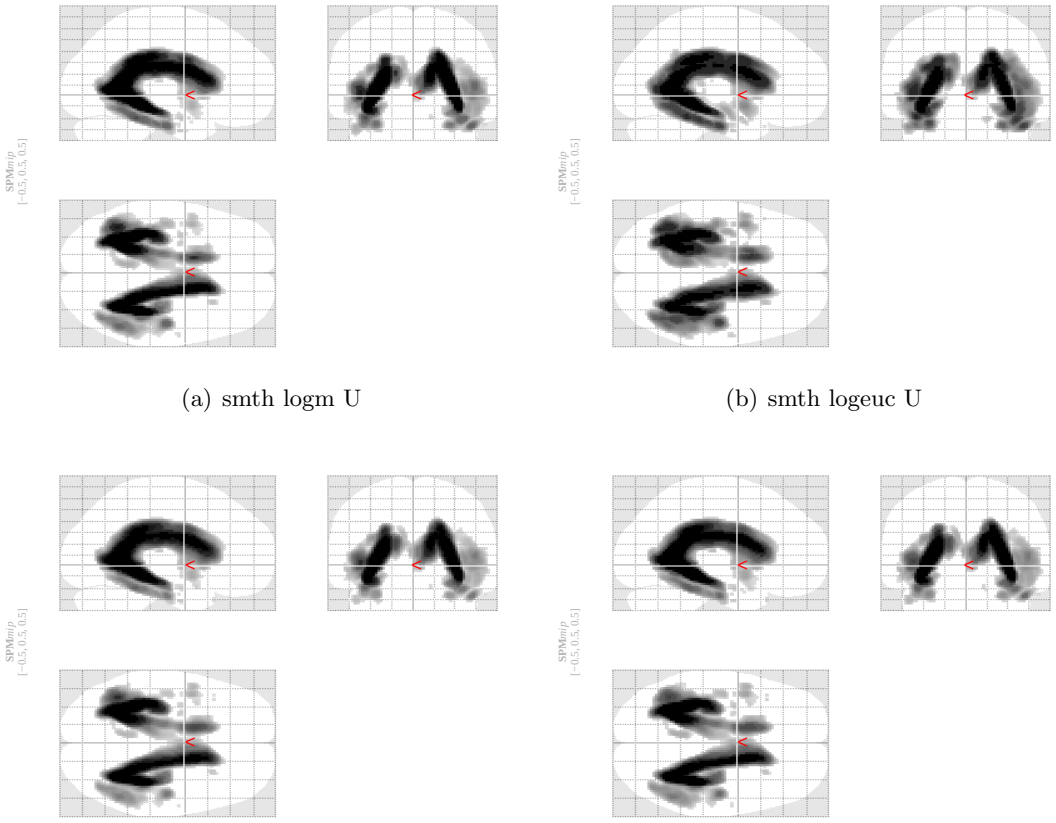
In the previous paragraph we have spoken loosely of log-Euclidean analysis. As discussed in the theory section of this chapter, there is a subtle difference between analysis of the unique elements of the Hencky tensor $\operatorname{logm}(U)$, and true log-Euclidean analysis of U, such that

$$\|\operatorname{vech}_{LE}(\operatorname{logm}(U))\| = \|\operatorname{logm}(U)\|_F.$$

We simultaneously address the closely related question of whether the large deformation tensors derived from $U = \left((I+K)^T(I+K)\right)^{1/2}$ are superior to the small deformation or infinitesimal strain tensor $F = (K^T + K)/2$ (analogous to the comparison of $\log|I+K|$ to $\operatorname{tr}(K)$).

Figure 4.30 compares MIPs for $pFWE < 0.05$ for Hencky tensor, the strict log-Euclidean version of U, the simple smoothed U, and the infinitesimal strain tensor F. Figure 4.31 quantifies the performance of the same TBM measures in terms of CDFs and voxel-wise comparisons, for uncorrected and for FWE-corrected p-values. The MIPs are almost identical, but a slight increase in significance is visible for the true log-Euclidean analysis. The p-value comparisons clearly favour the log-Euclidean approach; interestingly, this is one of the first such comparisons in this chapter for which the uncorrected and FWE-corrected results are in agreement. The CDFs and the voxel-matched comparisons are also consistent in favouring the log-Euclidean method. The other methods

framework, could help to improve sensitivity in multivariate DTI analyses, as it is known to be an important issue in scalar DTI-based studies [95].



(a) smth logm U

(b) smth logeuc U

(c) smth U

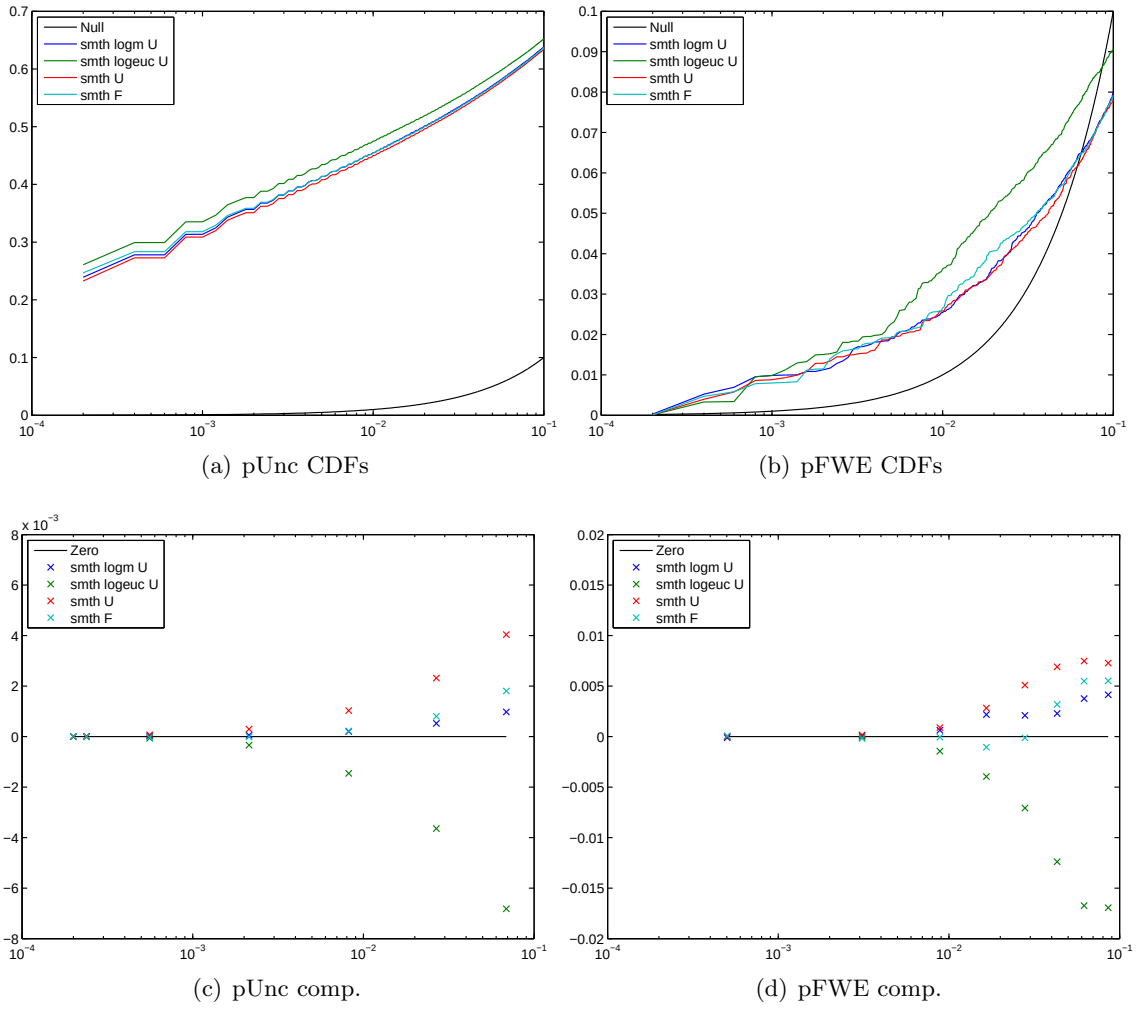
(d) smth F

Figure 4.30: Maximum intensity projections for different symmetric positive definite strain tensors. Large deformation tensors are based on $U = \left(J^T J\right)^{1/2}$, while the infinitesimal strain tensor is $F = \frac{K^T+K}{2}$. The distinction between 'logm' and 'logeuc' is that the latter scales the off-diagonal elements in the vectorisation so that the norm is preserved (it is $\text{vech}_{LE}(H)$ instead of $\text{vech}(H)$.

are not quite so clearly ordered, and, surprisingly, it appears that the infinitesimal strain tensor has produced very similar results to the finite strain Hencky tensor. It would be useful, in future work, to compare these tensors over a longitudinal interval longer than 12 months, and/or to compare them in a cross-sectional setting, to verify whether larger deformations confer greater benefit to the finite strain tensor.

Now we shift emphasis from the choice of TBM measure to more statistical issues. Firstly, we compare the Cramér statistic, which is suitable only for the simple two-group test performed here, with the more general Wilks' Λ statistic, which could be used for general MANCOVA designs. Figure 4.32 shows the statistic images, and the FDR and FWE p-values for two of the lower dimensional measures from table 4.4, while figure 4.33 shows the equivalent information for the two main higher dimensional options.

For the univariate determinant, the results are very similar, with limited evidence of greater power for the Cramér test. With the three eigenvalues, the Cramér statistic appears slightly superior in terms of FDR p-values, but quite dramatically better for FWE results. The full strain tensor and Jacobian matrix show very similar differences,



(a) pUnc CDFs (b) pFWE CDFs

(c) pUnc comp. (d) pFWE comp.

Figure 4.31: Comparison of tests using various strain tensors, using (left) uncorrected and (right) FWE-corrected p-values, in terms of (above) cumulative distribution functions, and (below) direct voxel-matched comparisons.

but with the nine-dimensional Jacobian leading to a more noticeable difference even in the FDR results. Note the important fact that Wilks' Λ exhibits (to an even greater extent) the phenomenon found earlier with the Cramér statistic: higher dimensional measures appear better in terms of FDR p-values, but indifferent or worse in terms of FWE; this is particularly noticeable in figure 4.32.

Figure 4.34 shows p-value CDFs for both statistics on all four TBM measures considered here. The Cramér statistic is universally superior when considering uncorrected p-values, and in terms of the FWE CDF, it is more powerful for all but the very strictest significance levels, below about 0.0004. Matched p-value comparisons (not shown) also favoured the Cramér test to Wilks' Λ for all four TBM measures, in terms of either uncorrected or FWE-corrected p-values.

Figure 4.35 illustrates the advantage of the step-down procedure for deriving FWE-corrected p-values from the distribution of the maximum-statistic. As expected, the general effect is an increase in power, with greater numbers of voxels being found significant at all levels from about 0.0005 to 0.2 (at which point the step-down procedure was ter-

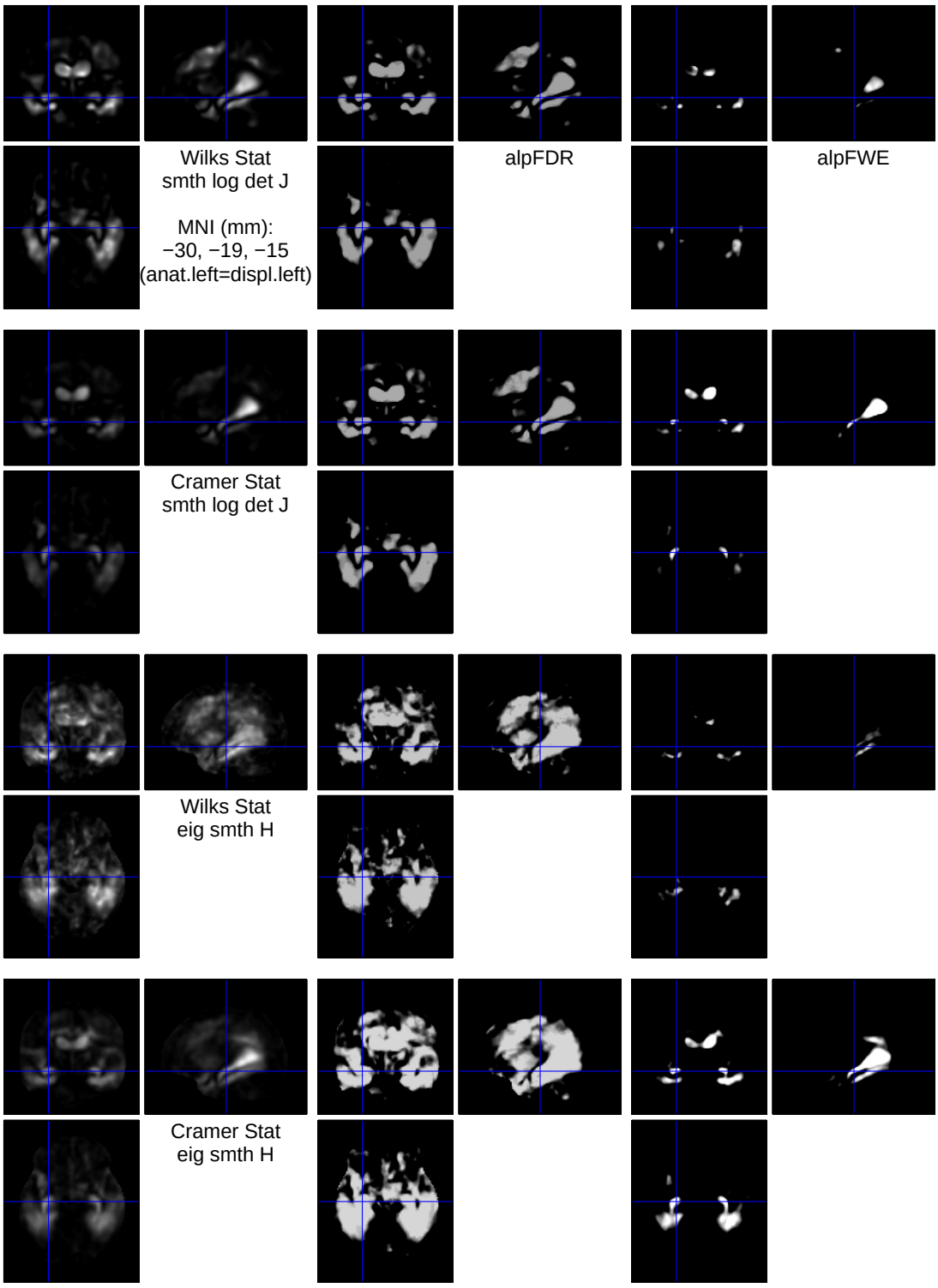


Figure 4.32: Statistical results for tensor-based morphometry, using the log determinant and the eigenvalues of the Hencky tensor, testing with either Wilks' Λ or the Cramér statistic. The Wilks-based statistic shown here is actually $1/\Lambda - 1$, which is zero for no effect, and larger for greater effects, as for the Cramér statistic. P-values are displayed in the range 0.05–0.0005 as absolute log p-values (brighter is more significant).

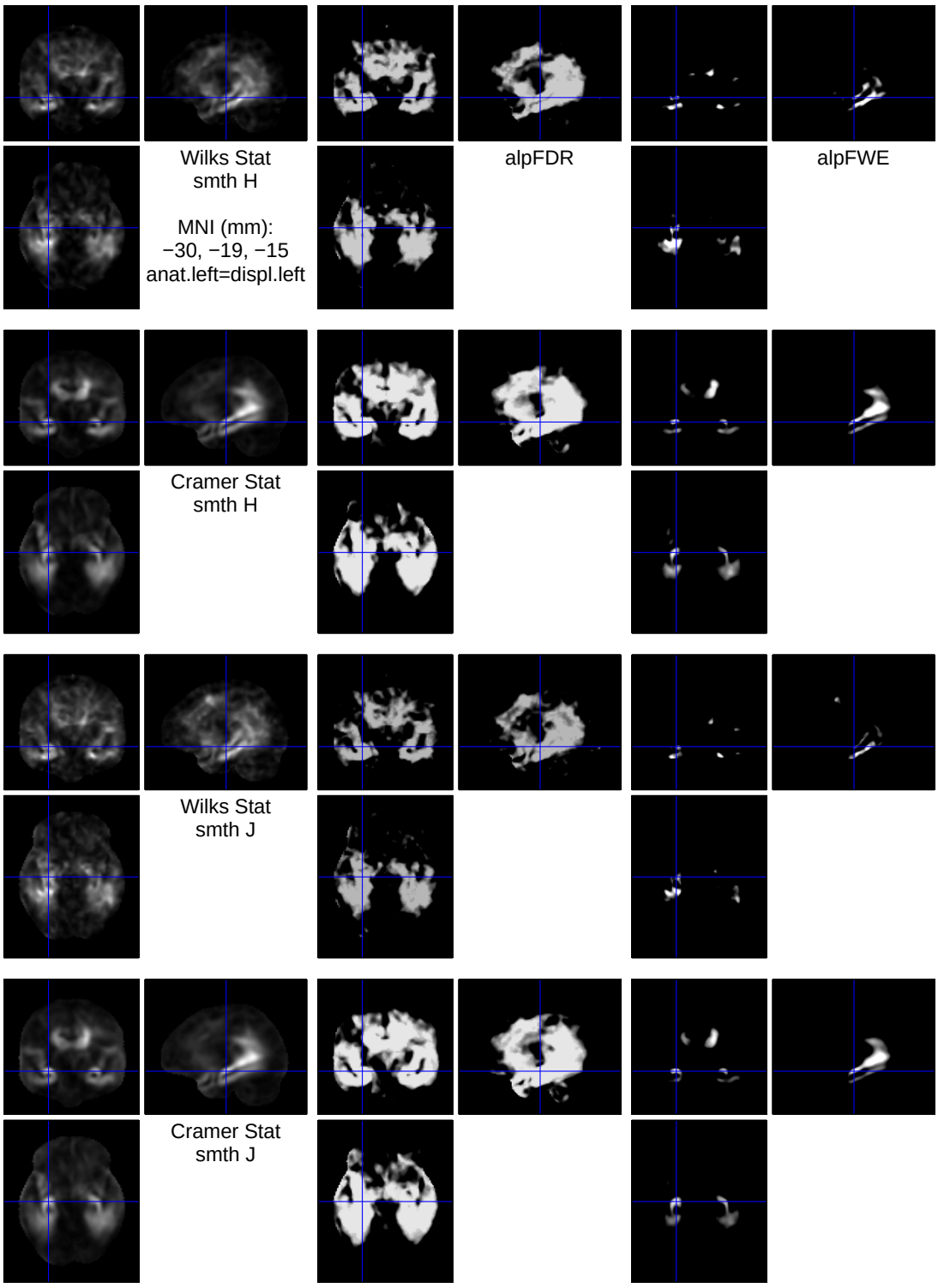
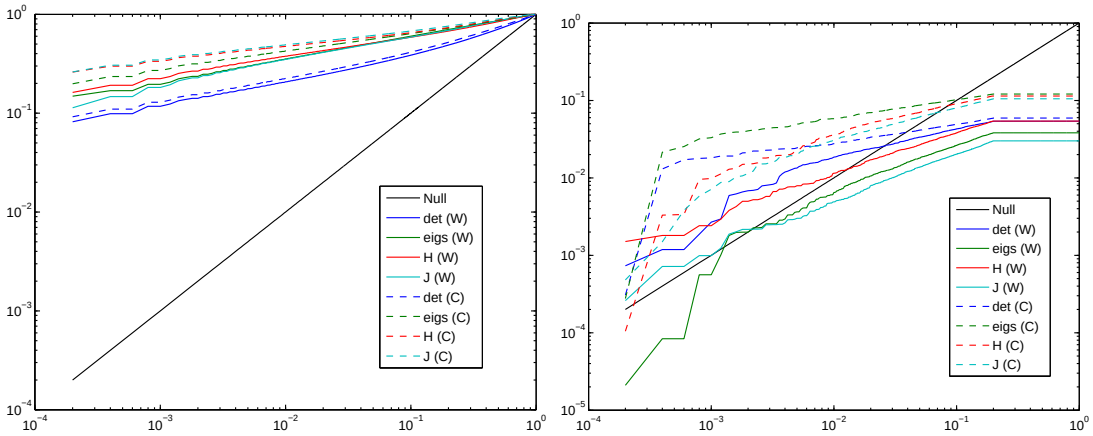
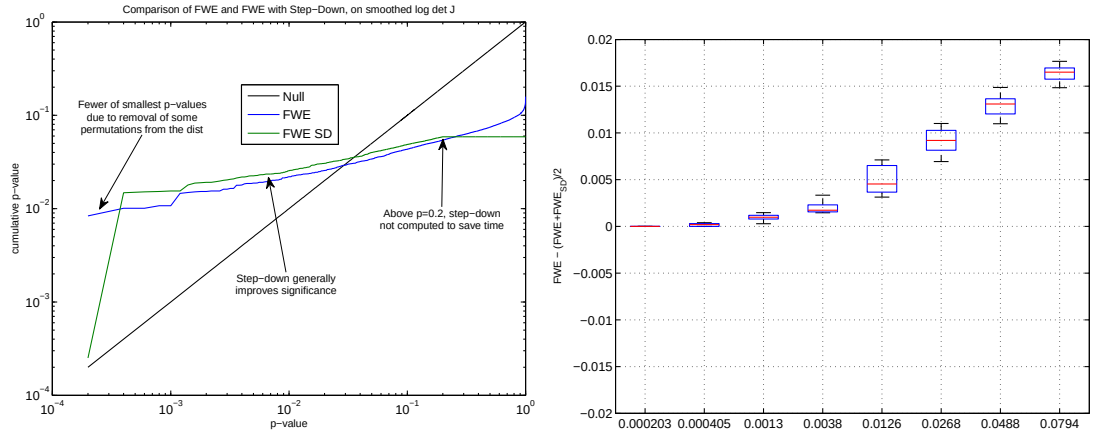


Figure 4.33: Statistical results for tensor-based morphometry, using the Hencky strain tensor and the full Jacobian matrix, testing with either Wilks' Λ or the Cramér statistic. The Wilks-based statistic shown here is actually $1/\Lambda - 1$, which is zero for no effect, and larger for greater effects, as for the Cramér statistic. P-values are displayed in the range 0.05–0.0005 as absolute log p-values (brighter is more significant).



(a) pUnc CDF (b) pFWE CDF

Figure 4.34: Comparison of tests using Wilks' Λ and Cramér statistics, in terms of (left) uncorrected and (right) FWE-corrected p-value cumulative distribution functions.



(a) Cumulative Distribution Function (b) Matched-voxel comparison

Figure 4.35: FWE-correction with and without the step-down procedure, for the smoothed log determinant. The higher values in (b) indicate superior significance for the step-down FWE p-values.

minated due to the fact that p-values over this threshold were considered uninteresting). For the very smallest p-values, between the minimal value of 1/5000 and about 5/5000, the step-down procedure reduces power. This would not occur with a true step-down algorithm, but is a consequence of our use of the Belmonte - Yurgelun-Todd approximation [92], which keeps track only of a certain number (12 here) of secondary-maxima to take the place of voxels removed during the step-down process, hence forcing a reduction in the effective number of permutations after all of the reserve voxels for a particular permutation have been removed. In this particular case, the first permutation to become exhausted occurs with the 73$^{\text{rd}}$ most significant voxel, which has the lowest possible p-value of 0.0002. The second exhausted permutation occurs with the 78$^{\text{th}}$ voxel, at the new lowest possible value of 1/4999. By the time the FWE p-value rises above 0.05, which occurred for the 9898$^{\text{th}}$ most significant voxel, a total of 550 permutations had been removed. The 14113$^{\text{th}}$ most significant voxel was the last to be considered below the arbitrary 0.2 cut-off, at

which point 4380 of the original 5000 permutations remained. The loss of over 10% of the permutations by the typical 5% alpha-level is slightly worrying, and perhaps suggests that for optimal results, either the number of reserves should be increased from 12, or the original total number of permutations should be increased from 5000 to maintain a reasonable number throughout the step-down procedure. However, the results here are more than adequate, as evidenced by the favourable comparisons of the step-down FWE p-values to the basic ones, in both (a) and (b) of figure 4.35.

Visualisation of orientational measures

Statistical results for generalised TBM on the set of orientational measurements discussed in section 4.2.9 are presented in the following subsection. First, illustrations of the measures for a single (quite severe) AD patient are displayed for the purpose of aiding visualisation and interpretation of the later findings.

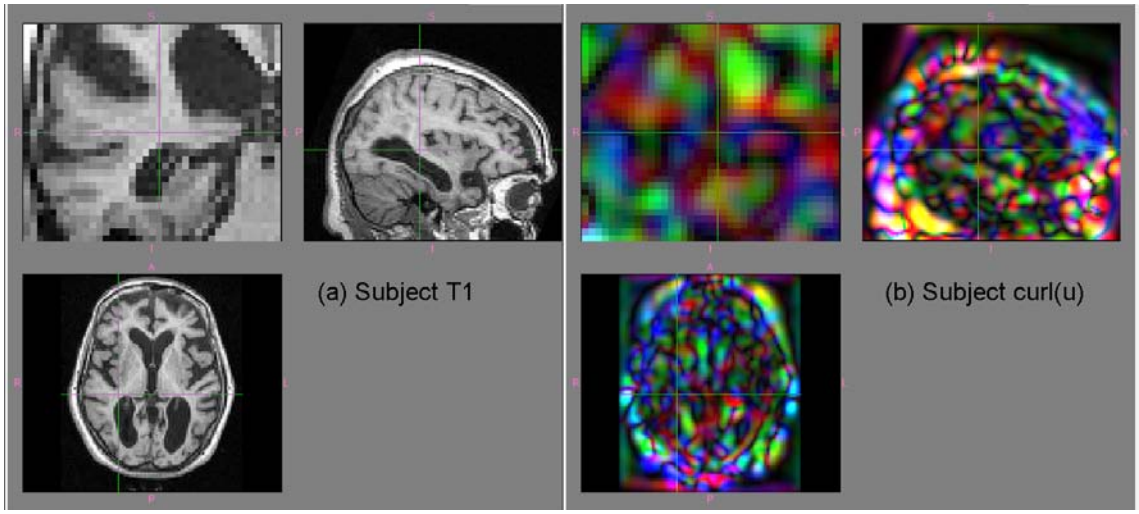


Figure 4.36: (a) Anatomical image, and (b) registered RGB colour overlay denoting direction (red — left/right , green — anterior/posterior, blue — superior/inferior), and magnitude (brightness) of curl. Anatomical-left is display-right. The cross-hairs are located at (33.5, -29.5, 2.5) mm MNI.

Figure 4.36 illustrates the orientation of the curl vector over the brain of a particular AD subject. For comparison, figure 4.37 shows the orientation of the displacement vector field. It appears that curl is a much noisier measure, as might be expected given that the processes of differentiation and of subtraction tend to amplify errors. However, it is still possible that information accumulated over multiple subjects could lead to significant levels of signal-to-noise; we later present statistic images and also compare group average images for curl, and for the other orientation measures illustrated next.

Figure 4.38 shows the (scalar) GA, for the same subject as the earlier figures. Note that in contrast to Diffusion Tensor studies, we would not necessarily expect clear differences between grey and white matter in terms of anisotropy of atrophy. Ideally, we would however see low GA in the CSF, since the fluid must in reality behave isotropically. It will be of interest to see if the tissue/fluid distinction is more pronounced in the groupwise

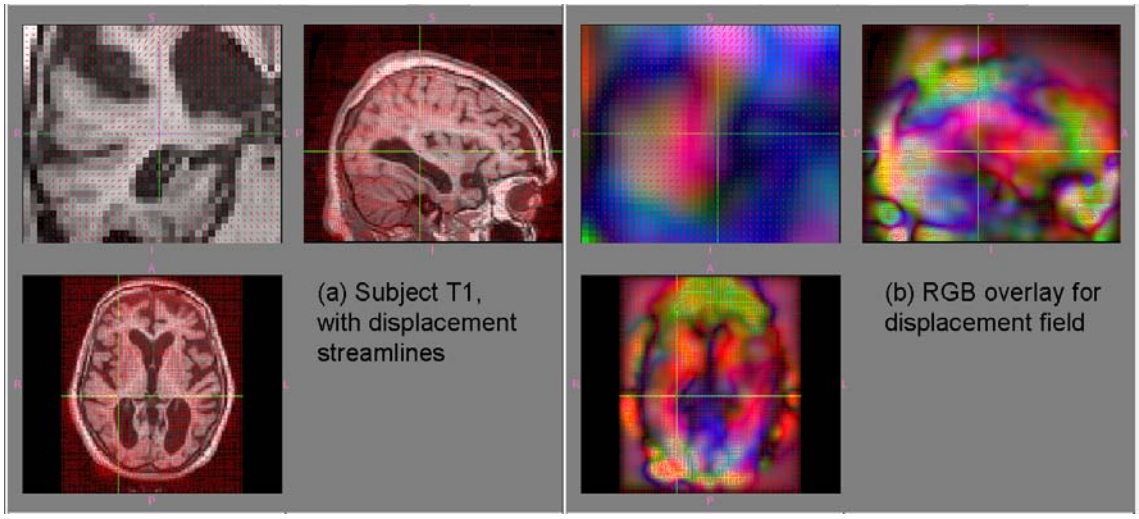


Figure 4.37: (a) Anatomical image with streamlines for displacement field, (b) corresponding RGB colour overlay (see fig.4.36). Anatomical-left is display-right.

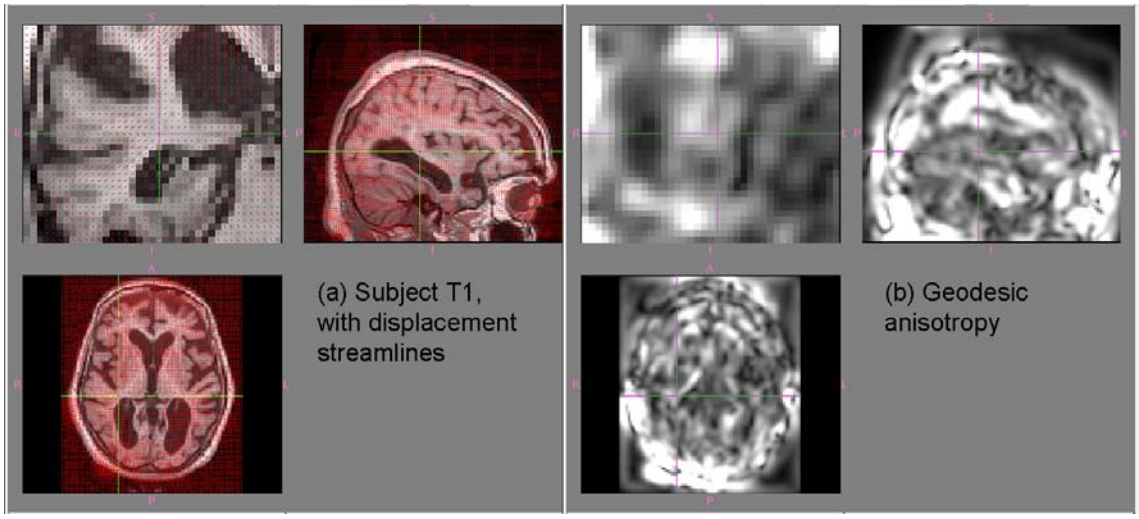


Figure 4.38: (a) Anatomical image with streamlines for displacement field, (b) Greyscale magnitude of log-Euclidean geodesic anisotropy. Anatomical-left is display-right.

results later.

Figure 4.40 shows separate x, y and z components for the principal direction, of the original (b) and smoothed (c) Hencky strain tensor. The components of the displacement field are shown for comparison. It is immediately obvious that the principal strain vectors are extremely noisy, compared to the displacement field shown in (a). Smoothing the tensor helps dramatically, and taking the absolute values further aids visualisation. Figure 4.39 shows an RGB overlay, which implicitly considers the absolute values.

In section 4.2.9 we suggested to analyse the vector obtained by scaling the principal direction by the principal strain and then taking the absolute values of the elements. Figure 4.41 illustrates the proposed measure in the same way as figure 4.40. It is surprising just how much more anatomically reasonable the results look, particularly when derived from the smoothed tensor (c). Note that the meaningful non-unit magnitude also justifies

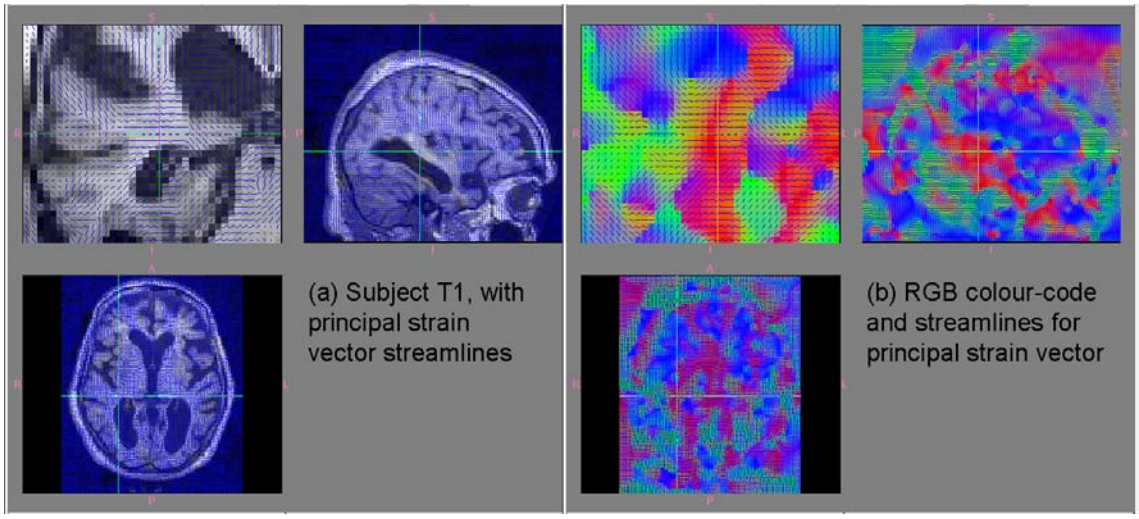


Figure 4.39: Illustration of the direction of the eigenvector of the Hencky tensor corresponding to the largest eigenvalue, (a) Subject T1, overlaid with streamlines, (b) RGB colour overlay and streamlines. Anatomical-left is display-right.

conventional smoothing of the three components when derived from an unsmoothed tensor, and this is shown in (d). Interestingly, smoothing the tensor seems to preserve more anatomical detail than smoothing the components, without an apparent trade-off in signal-to-noise. Figure 4.42 presents the view of this measurement corresponding to figure 4.39. Group-wise averages and statistical results for these two options should help to distinguish their relative merits.

Results for orientational measures

This section presents statistical results for the types of data discussed in section 4.2.9 and illustrated in the previous subsection. Because these orientational measurements are harder to interpret than simpler quantities related to displacement or volume change, we additionally present visualisations of the group-wise arithmetic means for each measure, similar to the single-subject visualisations given above. Results are shown for the control and patient averages superimposed on the single overall average T1 image.

First, to aid comparison, the (8 mm FWHM smoothed) displacement field averages are visualised in figure 4.43, at the same location and in the same way as the subsequent orientational results. The magnitude is generally larger for the patients, as expected, and is largest around the ventricles and the cortex, which is biologically plausible, but could also result from the limited T1-weighted intensity information present in bulk white matter away from tissue boundaries.

Figure 4.44 shows the magnitude and orientation of the group-average curl of the smoothed displacement field. Note that curl is a linear operator, so the group-wise averages of the curl of each subject's displacement field are the same as the curl of the group-wise average displacement fields shown in figure 4.43. The results are much harder to interpret, partly because the curl tends to have larger magnitude where the displacement magnitude is lower, suggesting that the displacement field has the potential to become more rotational

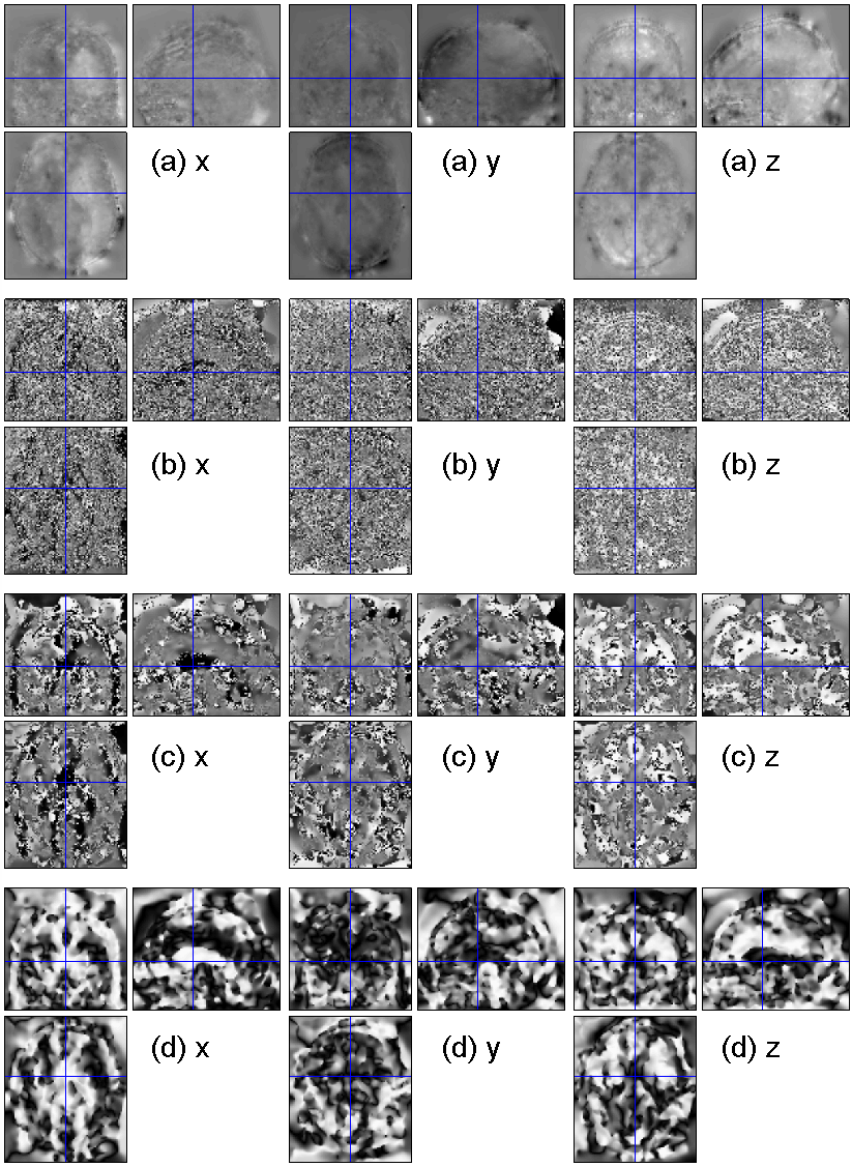


Figure 4.40: (a) Displacement field x, y and z components, (b) Principal strain direction components, (c) Principal direction from smoothed Hencky tensor, (d) Absolute values of images in (c). Anatomical-left is display-left.

when it is less constrained by the regularisation term; this also might indicate that the curl is more heavily (perhaps even predominantly) influenced by noise.

The geodesic anisotropy, visualised in figure 4.45 shows a more dramatic difference between control and patient means, with clear effects present in the hippocampus and temporal lobe. However, the finding of anisotropy differences within the ventricles is of course biologically implausible. This should not be surprising though, since the non-rigid registration software [55] was blind to the different tissue types, employing the same regularisation (based on logarithmic strains) everywhere in the field of view.

Figures 4.46 and 4.47 compare two of the options suggested in 4.2.9, based on the principal strain direction. Both measures show patient-control differences, though they are more pronounced and anatomically more clearly defined for the eigenvalue-scaled eigen-

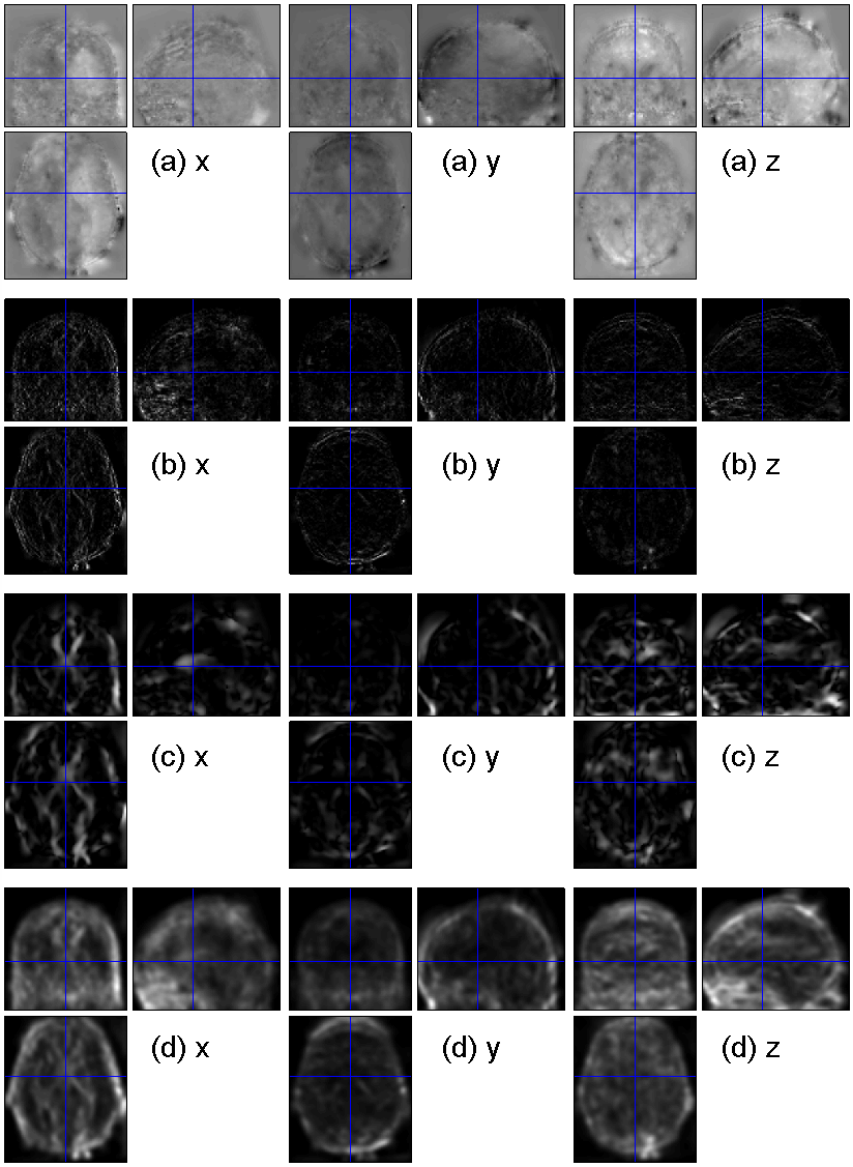


Figure 4.41: (a) Displacement field x, y and z components, (b) Absolute values of principal eigenvalue-scaled strain components, (c) As (b) but from smoothed Hencky tensor, (d) Smoothed components from (b). Anatomical-left is display-left.

vector shown in figure 4.47. The most easily interpretable difference (visible in the coronal view of the patient-mean, for both measures, but more so for the scaled one) is the vertical (blue) change in direction around the insula resulting from the opening up of the CSF space, characteristic of AD. Note that we have only visualised the results of the absolute scaled components of the principal vector from the smoothed Hencky-tensor, as in fig. 4.41(c), and not the results from smoothing the absolute scaled components from the unsmoothed tensor, as shown in fig. 4.41(d); the group-wise average results for the latter are very similar, but slightly less clear than those for the former.

Turning now to the statistical results, figure 4.48 shows images of the test statistic (Cramér or Watson) and thresholded maps of significance for five different orientational measures. Figure 4.49 shows corresponding maximum-intensity projections for the same

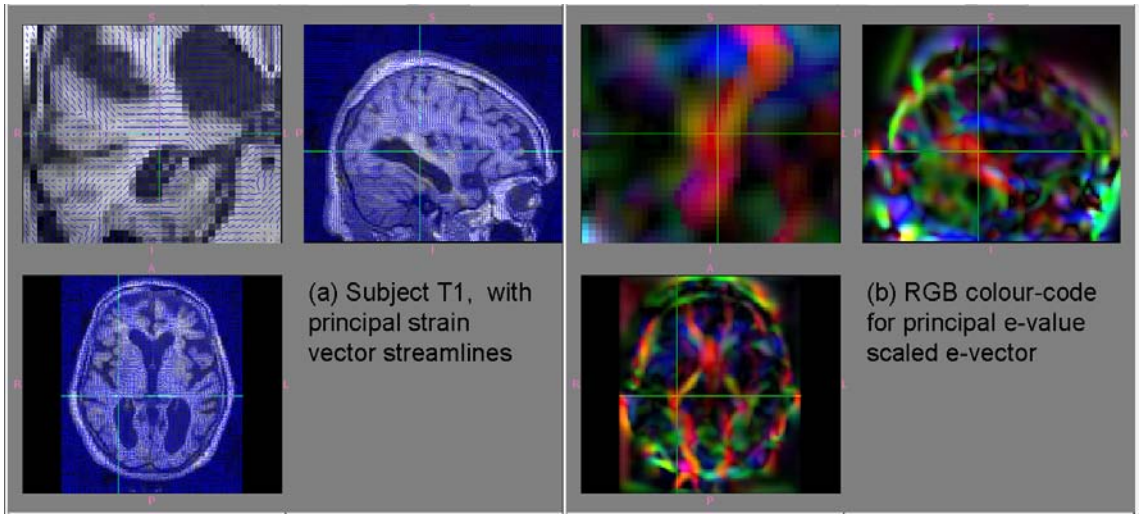
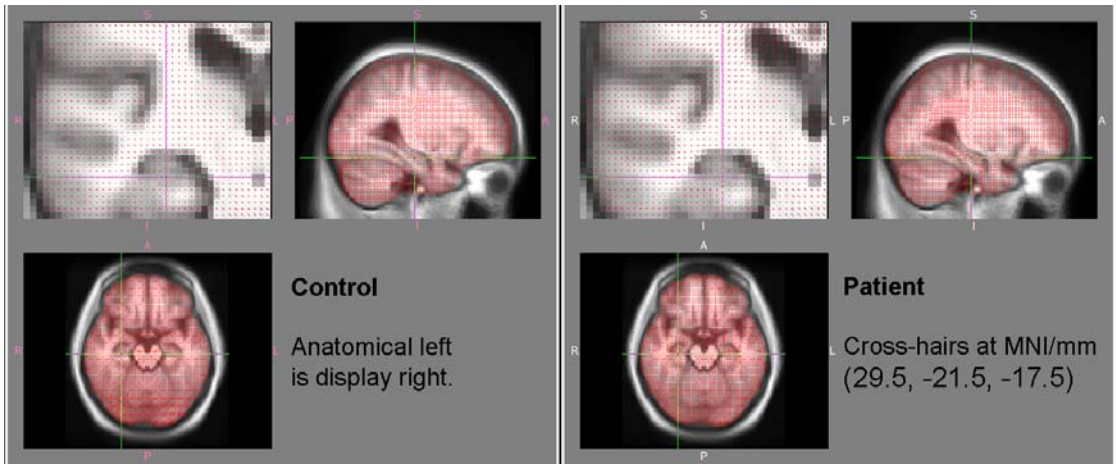


Figure 4.42: Illustration of the eigenvector of the Hencky tensor corresponding to the largest eigenvalue scaled by this eigenvalue, (a) Subject T1, overlaid with streamlines — same as fig. 4.39(a), (b) RGB colour overlay. Anatomical-left is display-right.

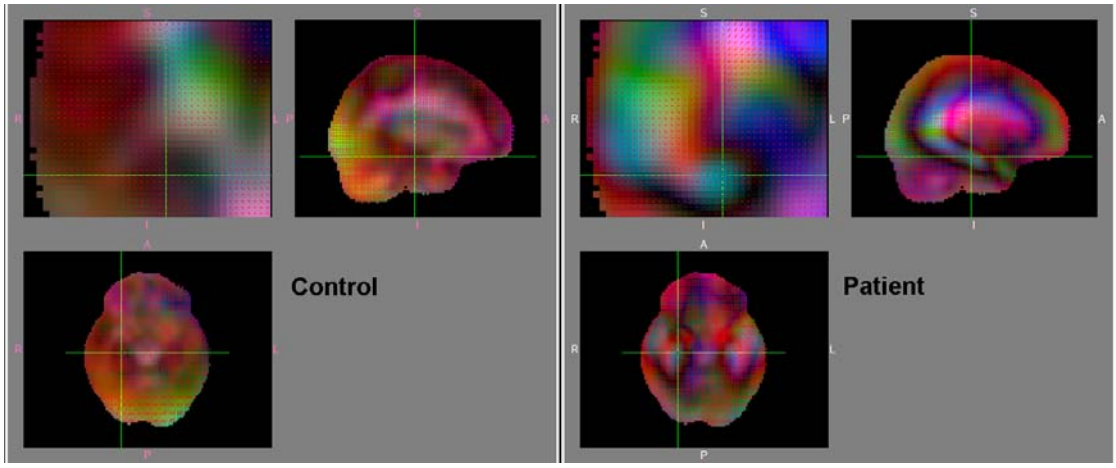
five orientational measures, with the standard log-determinant MIP provided for comparison.

One striking aspect from the MIPs is the greater roughness of the results from the unscaled principal eigenvector direction in fig. 4.49(c). This is understandable given the greater roughness apparent in this measure both in an individual subject (fig. 4.39) and in the group-wise averages (fig. 4.46). Corresponding results from (mis-)using the Cramér test on these directions (not shown) instead of the Watson test are rougher still, generally less significant, and anatomically harder to interpret. It appears that although this eigenvector comes from a smoothed strain tensor — just like the curl and GA which result in the smoother MIPs shown in figures 4.49(a) and 4.49(b) — the tensor smoothing is ineffective at spatially regularising its principal direction. A procedure for smoothing the unit vectors themselves, accounting for their manifold structure (i.e. not simply conventional smoothing followed by renormalisation) could be a useful topic for further research. Work on regularisation of diffusion direction maps [98] might be helpful in this respect.

The scaled eigenvector results shown in figures 4.49 (d) and (e) are much smoother; in particular, the former implies that blurring the tensor has a more pronounced smoothing effect on the principal eigenvalue than on its corresponding direction. Comparison of these two MIPs and of the results in rows 4 and 5 of figure 4.48 indicates a clear preference for the measure being derived from the smoothed tensor instead of smoothing the measure itself. Both scaled eigenvector measures yield results that are broadly similar to the volumetric log-determinant. This is unsurprising for several reasons, firstly, recall from equation 4.5 that the log-determinant is equal to $\mathrm{tr}\,(H)$ (in the absence of different smoothing or other preprocessing options) which is heavily influenced by the magnitude of the largest absolute eigenvalue of H that has been employed to scale the eigenvector direction. Also, taking the absolute values of the resultant scaled vector (with the intention of making it suitable for the Cramér test) further reduces the importance of the orientational aspect. In summary,



(a) Vector streamlines



(b) RGB overlay

Figure 4.43: (a) Anatomical average image with streamlines for group-wise averages of displacement vector fields, (b) corresponding RGB colour overlay denoting direction (red — left/right, green — anterior/posterior, blue — superior/inferior), and magnitude (brightness) of vectors.

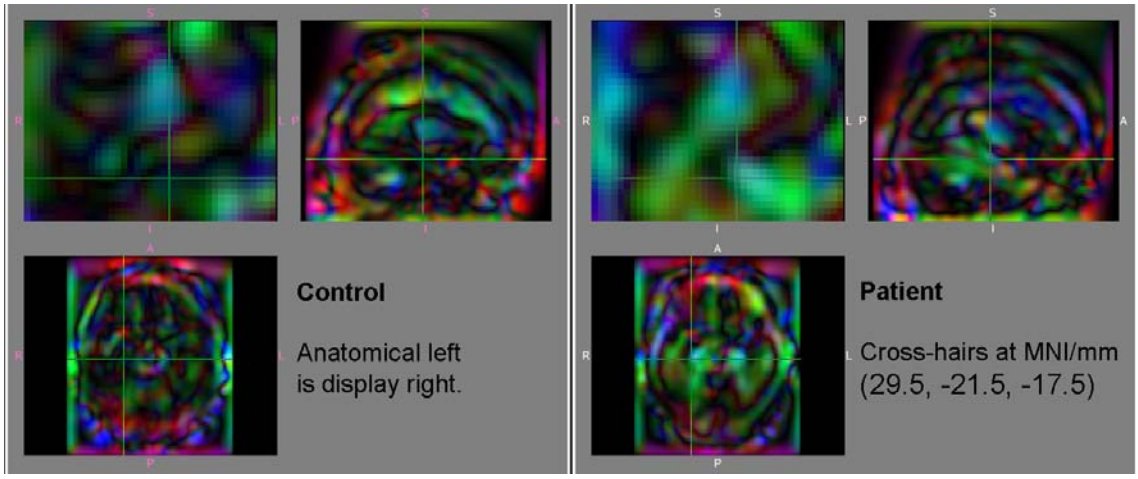


Figure 4.44: RGB colour overlay denoting direction and magnitude of group-wise averages of the curl of the displacement vector fields. See fig. 4.43(b) for key to colour-code.

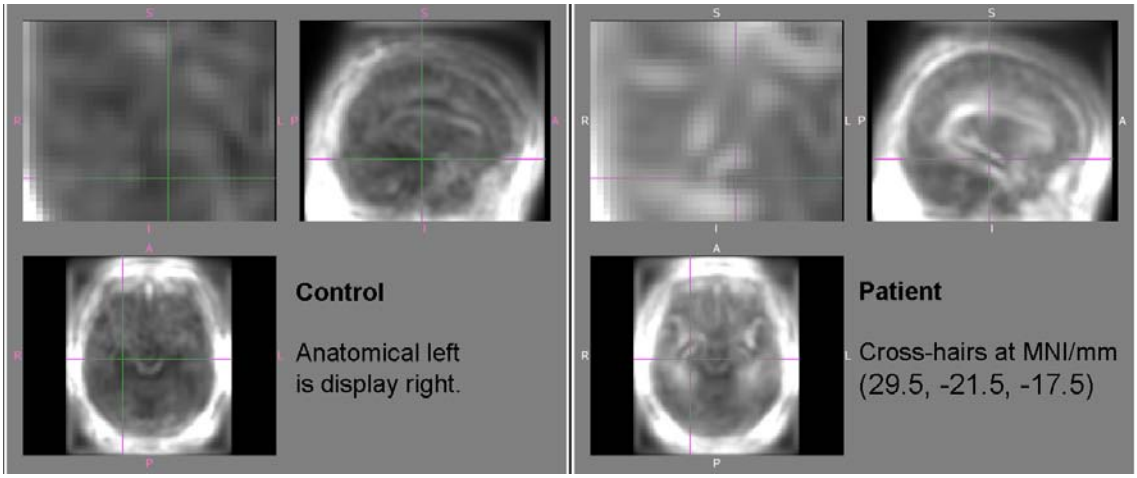


Figure 4.45: Greyscale magnitude visualisation of group-wise averages for geodesic anisotropy, derived from the smoothed Hencky tensor.

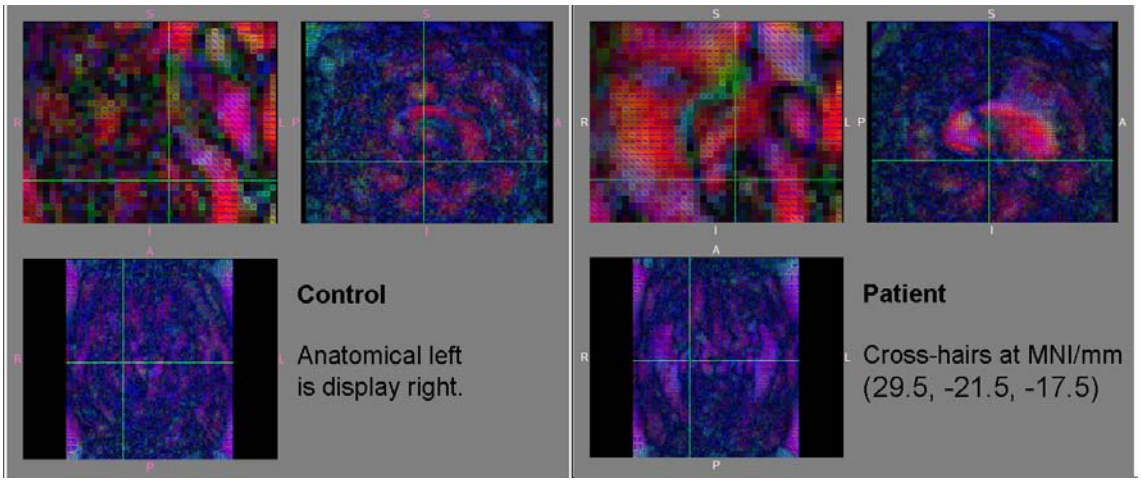


Figure 4.46: The group-wise mean direction from the eigenvector of the Hencky tensor corresponding to the largest eigenvalue, illustrated with an RGB colour overlay as in fig. 4.43(b).

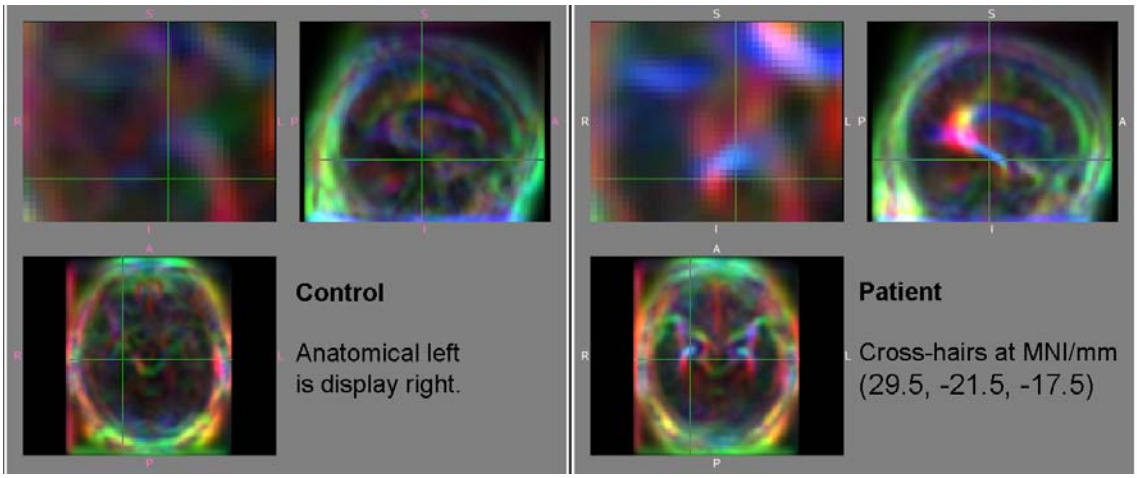


Figure 4.47: RGB overlay of the group-wise means from the eigenvector of the Hencky tensor corresponding to the largest eigenvalue scaled by this eigenvalue. See fig. 4.43(b) for key to colour-code.

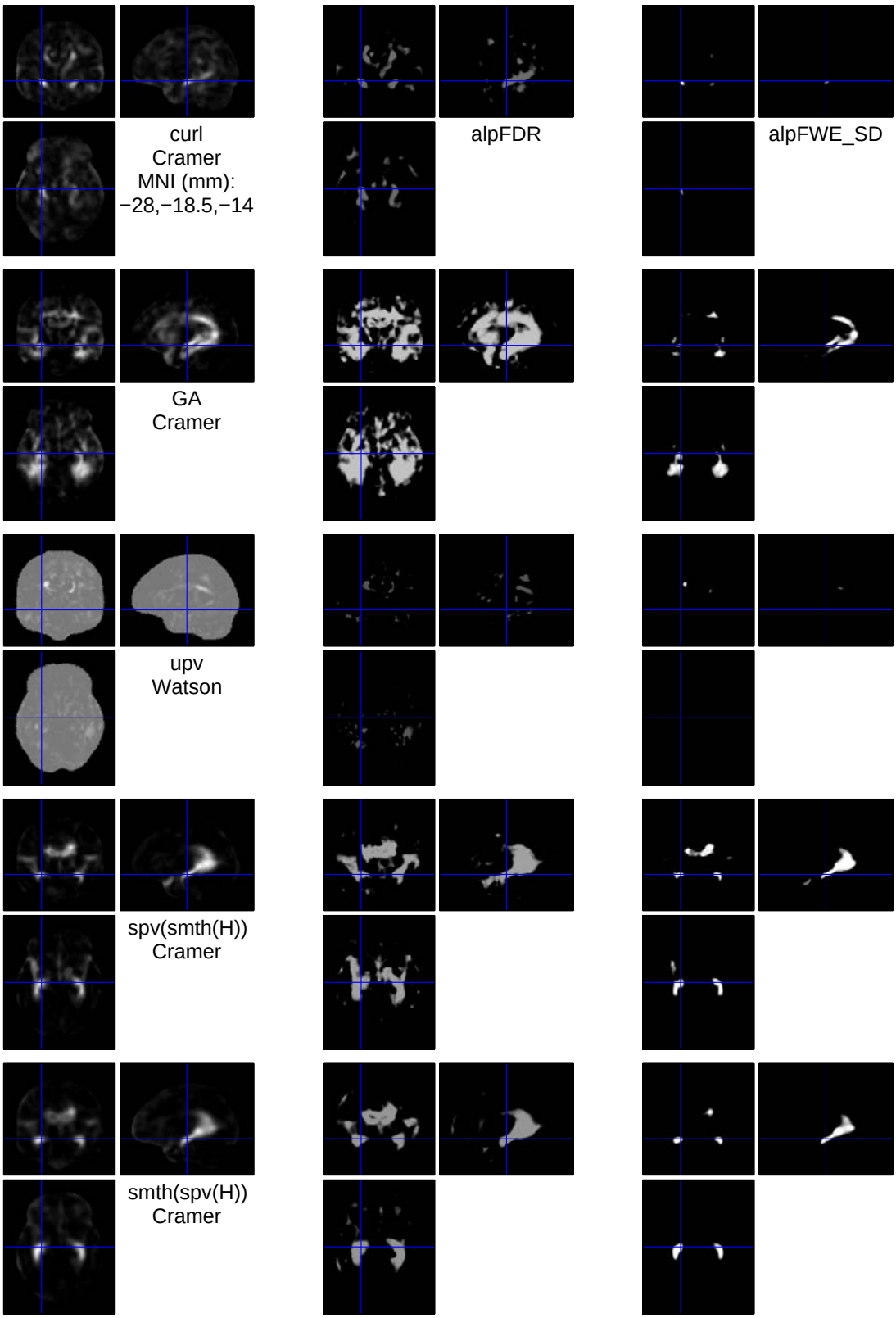
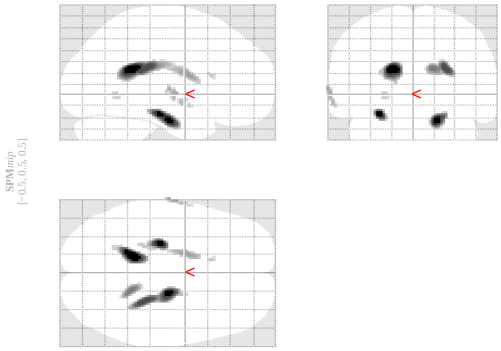
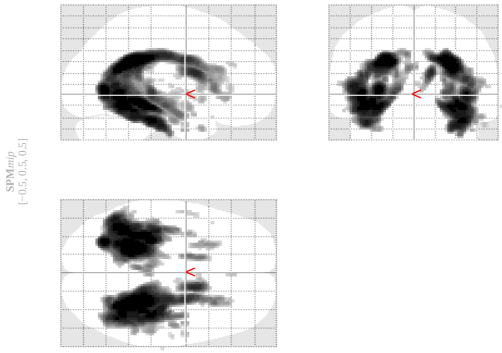


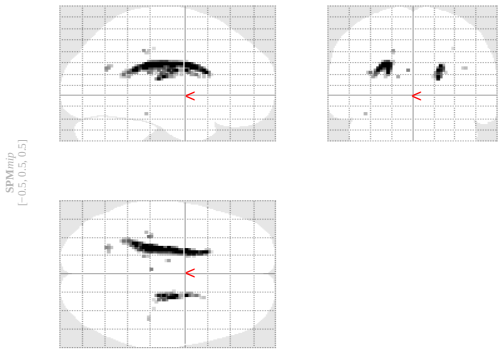
Figure 4.48: Statistical results for orientational measures. Rows, from top to bottom, are for: curl, geodesic anisotropy, unscaled principal eigenvector, eigenvalue-scaled principal eigenvector from smoothed Hencky tensor, and smoothed scaled principal eigenvector from unsmoothed H. Columns, from left to right, show: test statistic; FDR p-values; FWE p-values. P-values are displayed in the range 0.05–0.0005 as absolute log10 p-values (brighter is more significant). Anatomical-left is display-left.



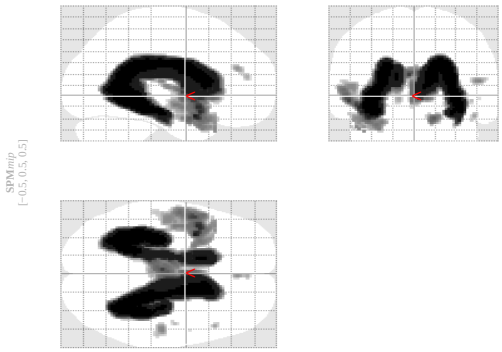
(a) Curl



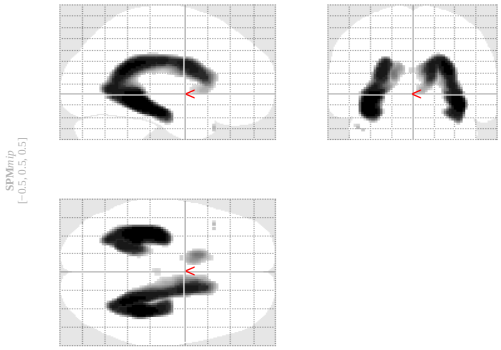
(b) GA



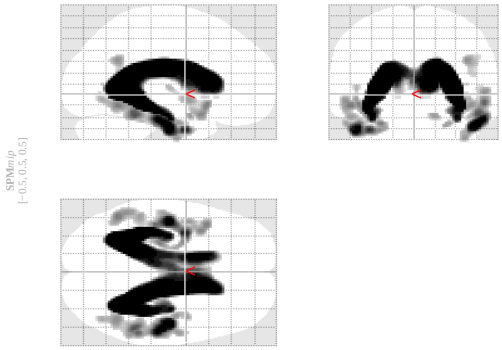
(c) upv (Watson)



(d) spv(smth(H))



(e) smth(spv(H))



(f) log det

Figure 4.49: Maximum-intensity projections of absolute log10 FWE p-values, thresholded at $pFWE < 0.05$, for the same orientational measures shown in figure 4.48 with the addition of the log-transformed Jacobian determinant to provide context. Anatomical-left is display-left.

introducing the scaling has brought greater spatial regularity and statistical sensitivity, but at the expense of the pure orientational interpretation and of some potential for the measure to complement the conventional volumetric one.

Returning briefly to the unscaled eigenvector direction, to address directly the question of complementarity, the results are in fact surprisingly disappointing; there are virtually no significant voxels from the Watson test of the eigenvector that are not also present in the Cramér test of the log-determinant. The removal of numerous voxels from the volumetric results might nevertheless be helpful in some circumstances due to the more precise interpretation afforded.

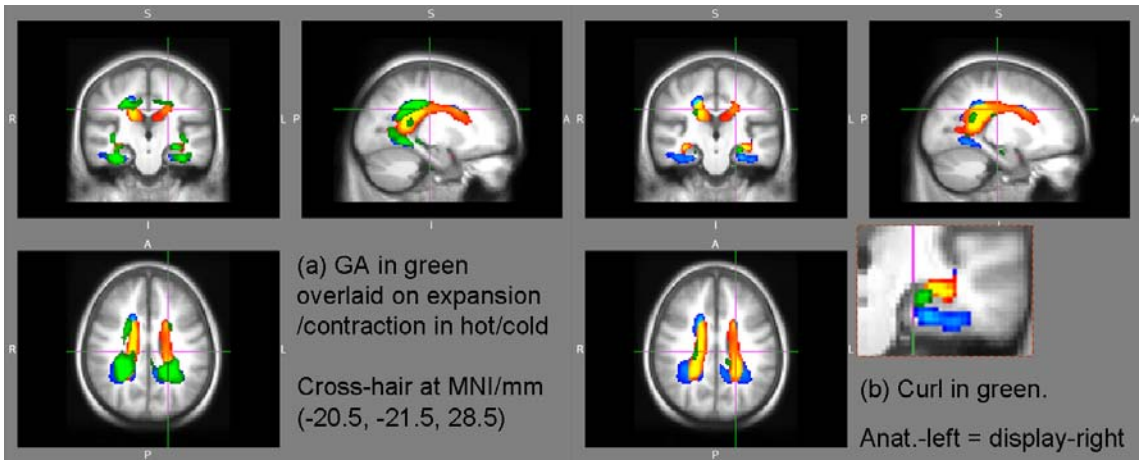


Figure 4.50: Results from the orientational measures of GA and curl overlaid on expansion/contraction coloured hot/cold (see also figure 4.54). All results are absolute log10 p-values for $pFWE < 0.05$.

The most exciting results from the orientational measures are those for the geodesic anisotropy and the curl. Figure 4.50 compares these to a volumetric measure (actually a similar, but more powerful strain measure). Unexpectedly, the GA is found to be more powerful in some regions than the strain-based measure,[36] adding a number of voxels in biologically plausible regions such as the insula and extending the regions of significance in the temporal lobes. The results for curl show relatively fewer additional significant voxels, but, interestingly, those which are added appear to be located with a high level of anatomically reasonable precision. As shown in the sagittal view and highlighted in the enlarged coronal image, significant vorticity has been found quite precisely in the gray matter of the hippocampal head. Intriguingly, reports of visually observed rotation of atrophying hippocampi have been made by clinicians at the Dementia Research Centre (private communication) which might be connected with the novel findings shown here, but further investigation would be needed to support stronger claims of association.

Figure 4.51 further investigates the extent to which different analyses might be complementary, focussing on the three measures proposed by Chung et al. [50], and expanded upon in section 4.2: the three components of the displacement field; its (scalar) divergence or volume dilatation; and its curl, equal to the three distinct elements of the in-

[36] Note though that the p-value CDF plots shown later (fig. 4.55) indicate that the full Hencky tensor (or its set of eigenvalues) are more powerful overall than the geodesic anisotropy.

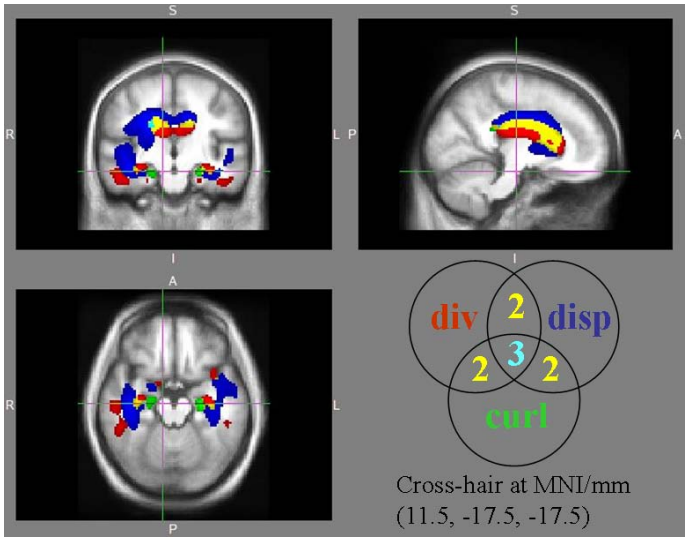


Figure 4.51: Venn diagram illustration of significant voxels ($pFWE < 0.05$) for the three measures proposed by Chung et al. [50]. (Anatomical-left is display-right.)

finitesimal rotation tensor. This figure is particularly noteworthy because Chung et al. themselves showed results only for displacement and dilatation, neglecting to present or discuss findings for curl. While the displacement field yields larger regions of significance, their anatomical interpretability is questionable, both in terms of the regions found, and, a priori, in the sense that they are expected to result from repositioning following volumetric changes. Vorticity suffers from the same difficulty that it may be driven largely by volumetric changes and associations with the regularisation method of the registration, but, at least in this case, the regions found are visually appealing. It is interesting to see how little spatial overlap is present between all three measures (cyan); this is an interesting practical analogue to the theoretical arguments in [50] that the measures should be statistically independent in terms of their random fields. Of course, statistical independence, a priori, is a distinct issue from spatial 'independence' of results, a posteriori; so it is not surprising that there are relatively large areas of overlap between two of the three measures (shown in yellow in figure 4.51).

The emphasis here is on the potential for orientational measures to find different patterns of significance to strain-based measures, in the hope of either locating additional areas, or of focussing the interpretation of common areas; the overall statistical power of the orientational measures is of less interest. Nevertheless, for completeness, figure 4.52 compares the observed sensitivities of the different orientational measures. The results serve mainly to reinforce earlier conclusions — that the geodesic anisotropy is the most sensitive of these measures, and that the scaled eigenvector is more powerful when the smoothing is performed on the tensor rather than the result.

Cross-methodological comparisons

We have thus far considered DBM and a wide range of TBM measures largely in isolation. In the comparison of orientational measures, figure 4.51 illustrated the complementary

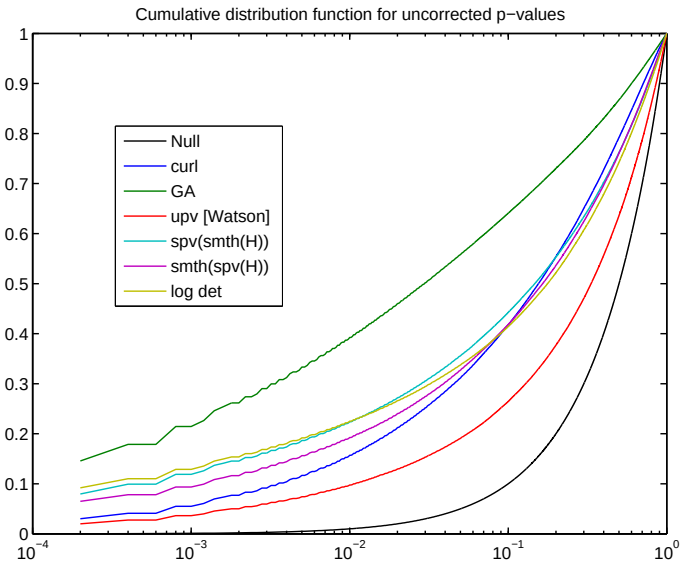


Figure 4.52: Statistical power of the different orientational measures illustrated via cumulative distribution functions of their uncorrected (permutation-based) p-values. The curve for log-determinant (which is similar to those for dilatation or the three components of the displacement) is shown for comparison.

nature of the curl and divergence of the displacement field with respect to the displacement vector field components. The purpose of this section is to briefly perform similar comparisons across the range of morphometry measures investigated here.

Figure 4.53 follows the same approach as figure 4.51, though the conclusions drawn from it are quite different. Whereas the orientational curl, and volumetric dilatation naturally contain different (and hence potentially complementary) information to the displacement, in this figure all three measures are quite closely related to strain. There are barely any voxels which are significant only for an individual measure; almost all of the voxels that are significant for the scalar measure are also present in both the higher dimensional measures, leading to a large core of cyan voxels. Furthermore, most of the voxels added to the classical measure by either of the generalised TBM measures are in fact common to both multivariate options, represented by the yellow areas.

An apparent disadvantage with multivariate generalised TBM measures, is that they lose the straightforward ability to interpret TBM findings in terms of expansion or contraction. However, we argue (novelly) here that there is a simple yet mathematically consistent way of providing this interpretation for several of the multivariate measures. First, we emphasise that this is not always straightforward: for example, it would not be reasonable to analyse the smoothed Biot tensor and then colour-code the results based on whether the smoothed determinant of J was above or below unity; the reason being that of the four operations involved, only the smoothing operation is linear, the inner-product and matrix square-root involved in $U = \left(J^T J\right)^{1/2}$ and the determinant operation are all nonlinear, and hence may not be simply interchanged with each other or with smoothing. In the theory section of this chapter, equation 4.5 shows that $\log|J| = \mathrm{tr}\left(\mathrm{logm}\left(U\right)\right)$, which is also clearly equal to the sum of the eigenvalues of $H = \mathrm{logm}\left(U\right)$. Therefore, if the

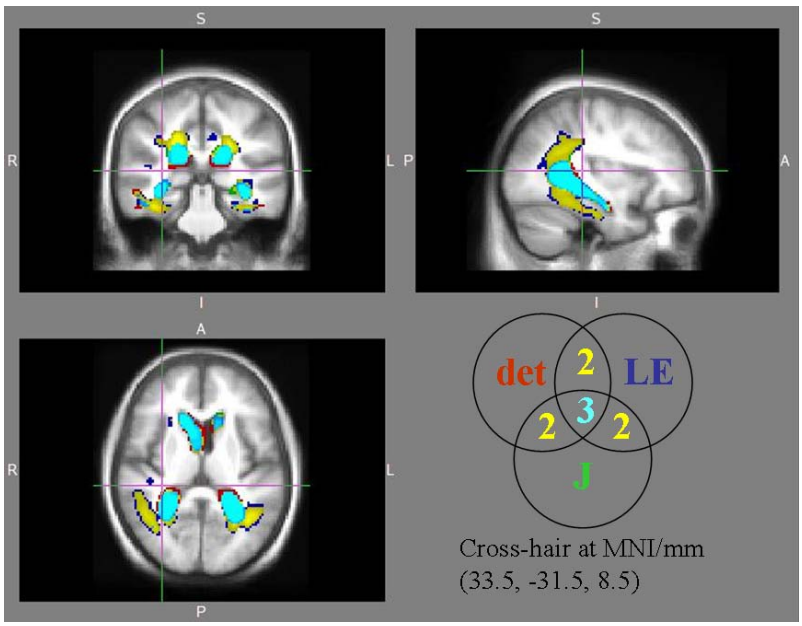


Figure 4.53: Venn diagram illustration of significant voxels ($pFWE < 0.05$) for the log-determinant, Log-Euclidean analysis of U, and the full Jacobian tensor. (Anatomical-left is display-right.)

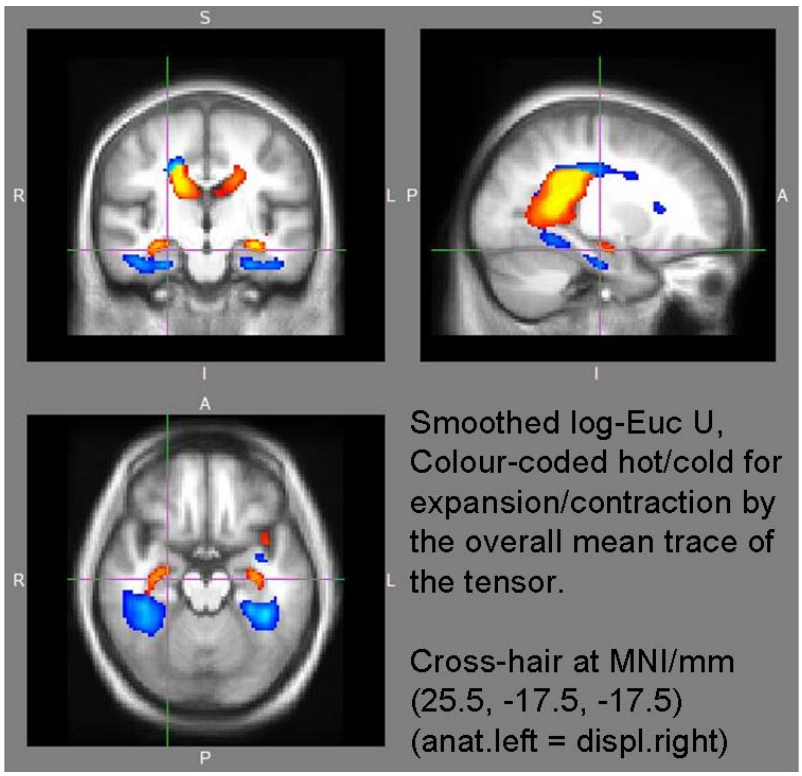
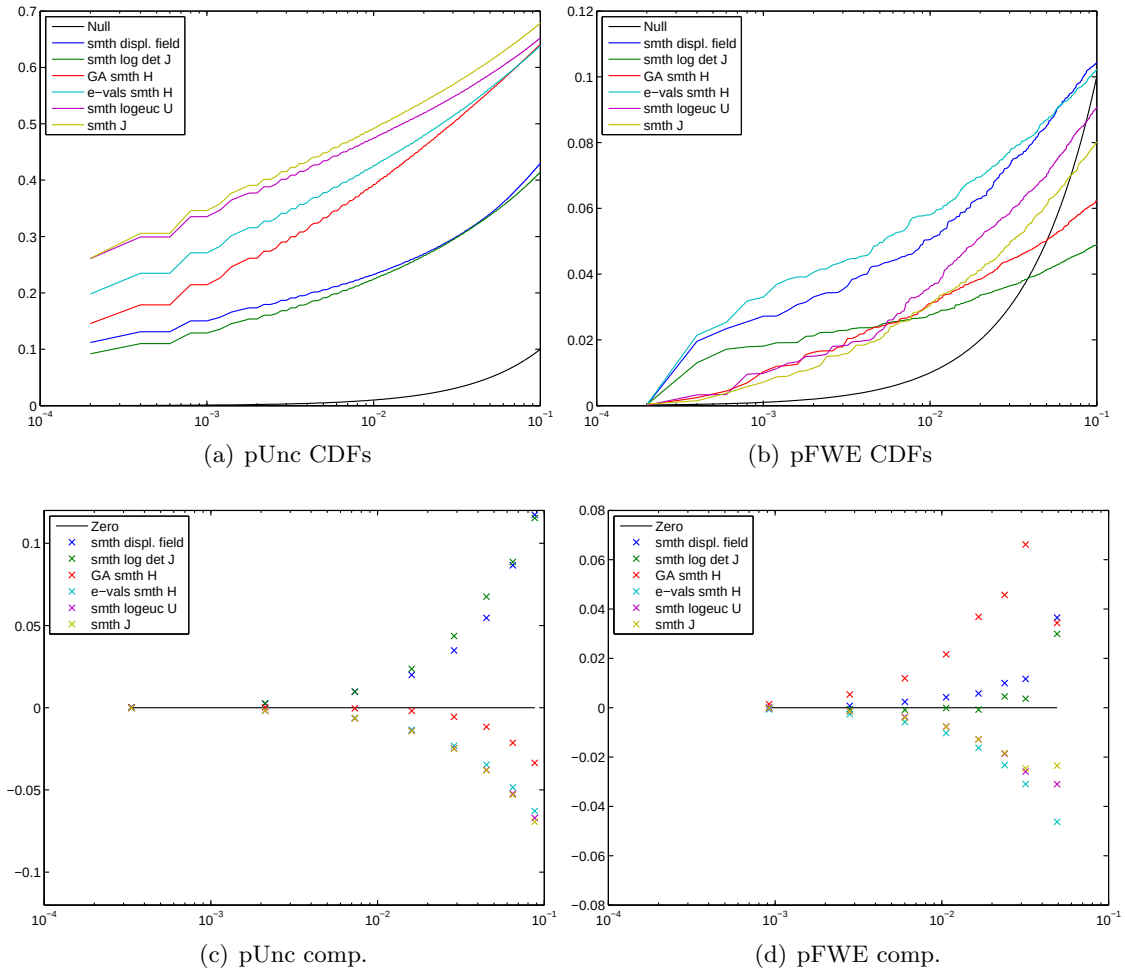


Figure 4.54: Significant voxels ($0.0005 < pFWE < 0.05$, on a log-scale) for Log-Euclidean analysis of U, colour-coded by the all-subject average map of $\mathrm{tr}\left(\mathrm{logm}\left(U\right)\right)$.

smoothing is applied to the Hencky tensor, one may *linearly* derive a quantity meaningfully related to $\log|J|$ (though note it will not be the same as the smoothed log-determinant) whose sign may then be used to colour-code analyses based linearly on H. Furthermore, the linear connection of the trace with the eigenvalues allows this powerful measure to be colour-coded by the sign of the sum of the eigenvalues. Equation 4.9 even shows that the geodesic anisotropy could be colour-coded in the same way, though this would not be easy to interpret. Figure 4.54 shows an example using $\mathrm{vech}_{LE}(\mathrm{logm}(U))$, which informatively distinguishes between expansion of the CSF spaces, contraction in the temporal lobes, and regions of presumably more complicated anisotropic strain sandwiched between expansion and contraction.



(a) pUnc CDFs (b) pFWE CDFs

(c) pUnc comp. (d) pFWE comp.

Figure 4.55: Comparison of deformation-based morphometry, tensor-based morphometry using log-determinant, and the geodesic anisotropy and eigenvalues of the Hencky strain tensor, using the Cramér test. P-value CDFs, and (below) matched voxel p-value comparisons; for (left) uncorrected and (right) FWE-corrected p-values.

Due to the similarity of the (non-orientational) Jacobian-derived measures, the key question becomes which of them is most powerful. Figure 4.55 presents CDFs and voxel-matched comparisons of uncorrected and FWE-corrected p-values. The superiority of the full Jacobian and the eigenvalues respectively for uncorrected and FWE-corrected p-values, as well as this discrepancy, has already been discussed; here, we focus on the relative merits

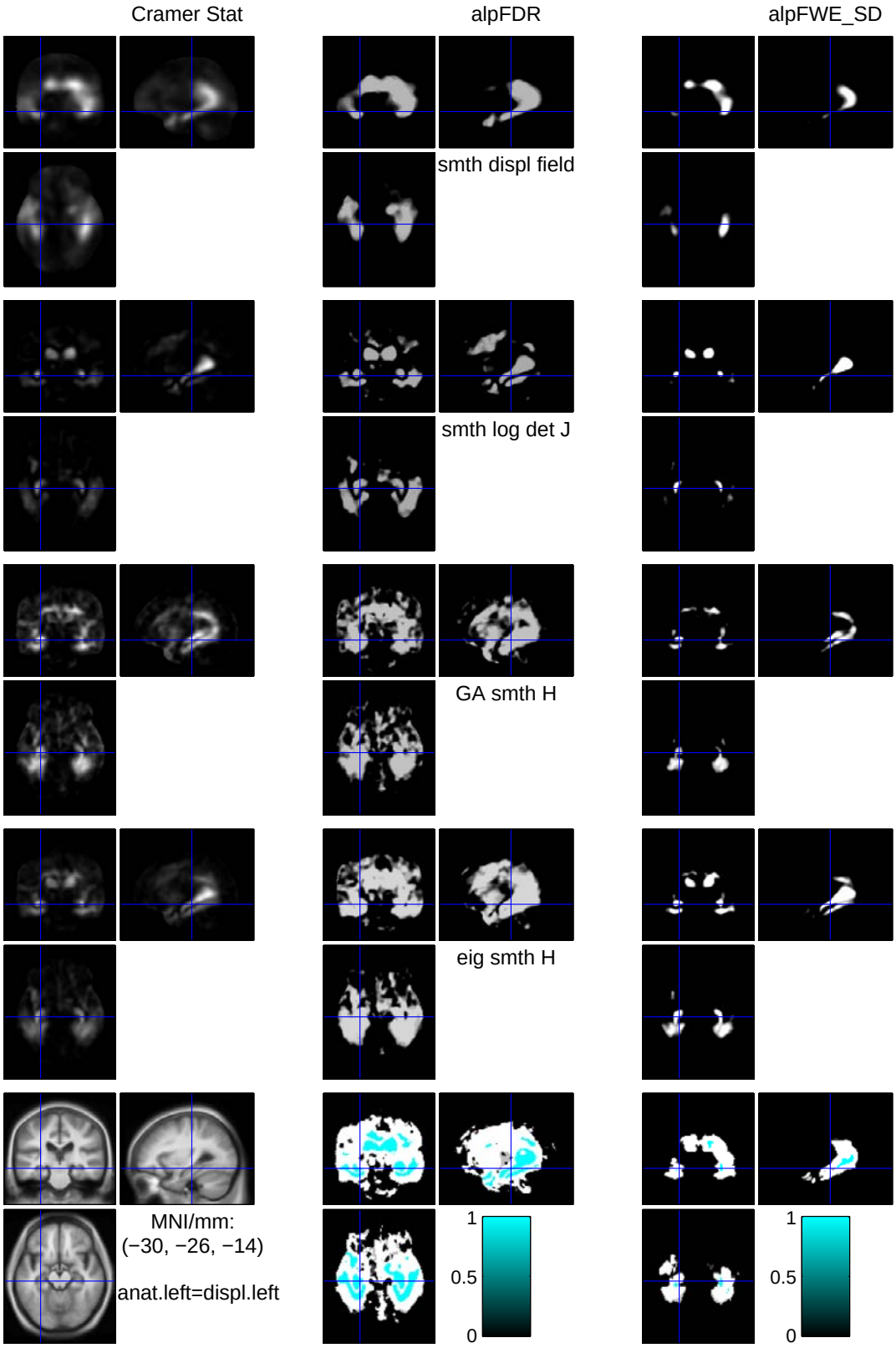


Figure 4.56: Comparison of deformation-based morphometry, tensor-based morphometry using log-determinant, and the geodesic anisotropy and eigenvalues of the Hencky strain tensor, using the Cramér test. P-values are displayed in the range 0.05–0.0005 as absolute log10 p-values (brighter is more significant). The final row shows the template, and Boolean unions overlaid with intersections (cyan) of the significant FDR and FWE results of the first four rows.

of DBM and the orientational GA among the other TBM measures. Uncorrected CDFs and voxel-matched comparisons show the GA to be slightly inferior to the eigenvalues, but clearly better than the log-determinant (the only measure here which shares its scalar dimensionality). The displacement field components perform marginally better than the log-determinant. Considering the corrected p-value CDFs, the displacement field moves up to be the second most powerful measure considered, and is closer to the eigenvalues than the tensors are to it. The geodesic anisotropy performs very similarly to H and J at strict thresholds, but is less powerful above about 0.002, it is below the determinant for levels stricter than about 0.0007 (as are the two tensors). For the voxel-matched comparison of pFWE, the displacement field performs worse than average, and the GA performs worst of all. However, this is more an indication of a problem with this kind of comparison than it is of disadvantages for these measures. The voxels considered are those for which the averages of all six measures' p-values were below 0.1 (and are sorted and binned based on this average). However, here, the displacement field and GA show quite different patterns to the four main TBM measures, and since there are twice as many of the latter, they dominate in the mean. If the p-value comparison was sorted by the displacement field, then it would probably appear to be one of the best measures. We nevertheless include this imperfect but useful summary, because it accounts for some spatial information missing from the CDFs (which are presented in isolation in [23]), since the latter wholly ignore the locations of the p-values, which are sorted independently for each measure.

Finally, figure 4.56 summarises the statistical findings for the cross-methodological comparisons considered here. The large differences between the union and overlaid intersection in the bottom row of the figure suggests once again that the alternative approaches may complement each other. In particular the displacement field has quite a different character (including, apparently, a difference in underlying smoothness, though it was smoothed with the same 8 mm FWHM Gaussian kernel) to the TBM measures, suggesting that it could usefully complement any analyses based on one or more of them. Interestingly, it appears that some, but not all, of the additional significance found for the eigenvalues compared to the log-determinant (equivalent to the sum of the eigenvalues in the absence of smoothing) can be explained as voxels with significant anisotropy. Since the eigenvalues have quite modest dimensionality compared with the Hencky or Jacobian tensors, and can be used to derive the GA and an optional measure of volumetric expansion or contraction, they seem to be the most appealing of the TBM measures considered in this chapter, and hence the strongest candidate for possible replacement of the standard log-determinant. The displacement field and its curl find different patterns of change, and hence are potentially important partners of generalised TBM using the Hencky eigenvalues.

4.5 Further work

We have mentioned several times through this chapter the need to evaluate the same methods again on different data-sets. However, now that the software and performance

quantitation techniques have been developed, this task could be performed as part of other work with a more applied nature. For example, clinical colleagues interested in the potentially greater power of generalised cf. standard TBM, might evaluate some of the different multivariate measures within a non-methodologically focussed paper.

One area where further technical development could be of great practical benefit to clinicians is in the realm of visualisation techniques. We attempted in section 4.4.4 to illustrate the orientational measures whose interpretation is particularly challenging, but there is undoubtedly room for improvement. We briefly mention the work of Wünsche et al. [53], who explore the use of a strain tensor (G in section 4.2.5) for the study of myocardial strain in a finite element model of the heart. Their paper presents numerous approaches for computer visualisation, including strain ellipsoids and 'hyperstreamlines', which could usefully be employed for visualising TBM data and results.

4.5.1 Diffeomorphic mappings

A particularly important topic for future research would be to investigate the potential for statistical analysis of diffeomorphisms (the theory of which was introduced briefly in section 1.5.1). The most accessible variant of these methods are those which produce diffeomorphic transformations via exponentiation or integration of stationary velocity vector fields [99]. This corresponds to analysing one-parameter subgroups of diffeomorphisms in their tangent space at the identity, and could be achieved relatively easily in practice thanks to the availability (in SPM5) of Ashburner's fast DARTEL algorithm [37].

There is scope for both high-dimensional multivariate analysis of complete velocity fields and for local voxel-wise analysis of the vectors, or measures derived from them, in place of the products of the resultant displacement vector fields analysed here. Arsigny et al. [100] provide algorithms for computing the logarithmic map (from a deformation to the velocity field or tangent space) as well as the exponential map (which corresponds to integrating the velocity field); these could potentially be used to derive diffeomorphism-group representations for analysis from conventionally computed deformation fields.

As mentioned in [37], momentum maps from the more mathematically sophisticated diffeomorphic metric mapping framework [101, 102, 103] could provide a spatially sparser representation of large deformations, potentially better suited to statistical analysis (particularly classification approaches [104]). These methods allow more general diffeomorphisms (not contained in one-parameter subgroups) that correspond to integrating a time-varying velocity field, though Hernandez et al. [99] found only a very minor improvement in accuracy from this significantly more computationally-demanding setting.

4.5.2 Alternative statistical methods

A potentially powerful extension would be the consideration of cluster extent within the multivariate morphometry framework; for example using the permutation distribution of cluster size [93] or mass [105, 106] to provide FWE corrected inferences that favour larger connected components of morphometric change. Alternatively, voxel intensity and cluster extent information could be fused using the combining function approach of Hayasaka and

Nichols [107]. For larger multivariate observations (e.g. the full Jacobian matrix) at high resolutions, these techniques will be very computationally demanding; particularly since the need to determine the size of connected components complicates the use of memory-efficient blocking strategies (appendix D.3), because the connected components could cross over the predefined block boundaries.

The searchlight, smoothing, and Bayesian methods

The searchlight technique [1] investigated here is one attempt to improve upon the shortcomings of conventional spatial smoothing. Its principal advance is that by replacing simple weighted-averaging by multivariate analysis of the data, it should have the potential to detect more complex high-resolution patterns of effect. However, it has not performed particularly well on our morphometric data, in comparison to simple Gaussian smoothing. The most notable limitations of the searchlight method, with particular relevance to inter-subject studies that are almost universally performed for structural imaging, are that it remains stationary, isotropic, and purely distance-based.

Recent Bayesian formulations of statistical parametric mapping techniques [108, 109, 110] have overcome some of the limitations with smoothing in classical SPM analyses, and may also represent an improvement over the searchlight method. In particular, by including a model of the signal's spatial regularity (in subtle distinction with the standard attempt to smooth away the noise) it becomes possible to adaptively learn the smoothness from the data itself [111, 112], to separately model signal and noise process regularity [113], and even to estimate a non-stationary local smoothness [114, 115]. Perhaps most impressively, recent work by Harrison et al. [116, 117, 118] allows Bayesian estimation of non-stationary anisotropic smoothness, similar to early work on anisotropic filtering [119], but replacing arbitrary preprocessing decisions with principled Bayesian model comparison (e.g. using the Bayesian evidence framework to infer that a particular fMRI study is better modelled with non-stationary processes [117]). This exciting work has yet to be applied to inter-subject analyses or structural data, to the best of our knowledge, but it seems likely that it could contribute dramatically to morphometry.

Novel measures and tests

We now propose the use of some new TBM options, which have not, to the best of our knowledge, been mentioned elsewhere.

Note that the fractional anisotropy can be rewritten as the magnitude (Frobenius norm) of a normalised 'deviatoric tensor' [62]

$$FA = \frac{\sqrt{3}}{\sqrt{2}} \frac{\|D - I\bar{\lambda}\|_F}{\|D\|_F} = \left\| \frac{D - I\bar{\lambda}}{\|D\|_F \sqrt{2/3}} \right\|_F$$

where $\bar{\lambda} = \mathrm{tr}\,(D)/3$ is the mean eigenvalue. Similarly, the geodesic anisotropy can be expressed as the magnitude of $\mathrm{logm}\,(D) - I\mathrm{tr}\,(\mathrm{logm}\,(D))/3$ — essentially an unnormalised equivalent on a logarithmic deviatoric tensor. A natural extension from univariate analysis

of FA or GA would therefore be to consider multivariate statistical analysis of all six unique components of the (normalised or logarithmic) deviatoric tensors. This offers no dimensionality reduction over analysis of the original tensors themselves, but may allow more precise interpretation of significant findings stemming from differences in anisotropy instead of differences in the tensor eigenvalue magnitudes or eigenvector orientations. This could be usefully applied in both TBM and diffusion tensor imaging.

Regarding testing of general (SPD) tensors, in the special case of a two-sample design, we have used the Cramér test with Euclidean distances, as it is was originally defined, and also with (effectively) log-Euclidean distances,[37] as was done by Whitcher et al. [16]. On a related, but distinct point, it was outlined in section 4.3.3 that that the Euclidean distances can be subsequently transformed through the use of a kernel function, with the aim of sensitising the test to a particular form of alternative hypothesis.

Surprisingly, it would appear that we are the first to recognise that the test seems only to depend on the concept of inter-point distances; not specifically on Euclidean inter-point distances. This means that in addition to simple preprocessing of the data, or postprocessing of the distances, it should also be possible (in fact quite simple) to base the test directly on a Riemannian distance metric. This is a particularly appealing combination; firstly, the full affine-invariant tensor distance [58, 60] promises superior theoretical properties. Secondly, the use of a Riemannian Cramér test brings an additional practical benefit: within the framework for Riemannian analysis proposed in the literature to date [58, 60] it is necessary to iteratively compute an estimate of the mean — an expensive procedure when it must be done at every voxel; however, the Cramér test requires only the Riemannian inter-point distances themselves, not the mean, and these can be computed in closed form.

This novel Riemannian Cramér test also offers an alternative to the Watson test for testing principal strain (or diffusion) axes. The test presented in section 4.3.3 from Schwartzman et al. [64], computes a 'mean' direction and estimates of 'dispersion' based on the Bipolar Watson distribution, and involving eigenvalues of various matrices. The dispersions are then used to construct an approximate F-statistic (which we test with permutation, avoiding the need for parametric assumptions). It is not clear what, if any, Riemannian metric has this mean as its geodesic barycentre, nor whether the dispersion relates to any geodesic measure of distance between axes. Excitingly, there is a trivially simple natural Riemannian distance between axes, simply given by the angle between them (e.g. computed from the inverse-cosine of the dot-product of the unit vectors); we are not aware that it leads to any simple concept of a mean direction (for more than two axes), but it would seem to be enough to base a two-sample Cramér test upon. In future work, we will directly compare log-Euclidean and affine-invariant Cramér testing of strain tensors, and the Watson and our new Riemannian Cramér test on principal axes of strain.

[37] The distance used was actually still Euclidean, but the data had been log-transformed in advance.

4.6 Conclusions

In this chapter, we have explored several multivariate generalisations of deformation- and tensor-based morphometry. We have presented original results for searchlight DBM and TBM, and the first application to morphometry of some orientational measures proposed for diffusion tensor imaging. Capitalising on the work in chapter 2, we derive the first family-wise error corrected results for the Cramér and Watson statistical tests. We have shown that the Cramér statistic outperforms the more general Wilks' Λ statistic, in the two-sample situation for which it is appropriate, particularly when the dimensionality of the data causes the number of unique covariance matrix elements to approach the number of observations. We have also suggested a novel Riemannian formulation of the Cramér test, which we believe could have great potential in both generalised TBM and DTI analysis.

Some limitations must be admitted, the gravest of which is probably our use of low-dimensional DCT-based spatial normalisation instead of modern high-dimensional group-wise registration methods [37, 48, 120]. The resulting lack of precise spatial correspondence,[38] is also the cause of the second major limitation, which is our use of a relatively large spatial smoothing kernel. We have ameliorated this issue to some extent through our consideration of the searchlight technique, though it is likely that more precise registration would result in better performance of the searchlight, as well as decreasing the size of the optimal smoothing kernel.

Regarding the discrepancy between uncorrected/FDR and FWE performance of the multivariate measures, the failure of either smoothness or the maximum distribution to account for this motivates further research. It will be important to see if this phenomenon is replicated using other data-sets (though note that we have already replicated it using the searchlight technique in section 4.4.3 and using generalised TBM over both the 12 and 6 month intervals). In addition to analysing further real data-sets, it would be particularly helpful to investigate simulated data. For example, one could sample multiple Gaussian random fields with different levels of spatial correlation and explore how both the smoothness and an increasing number of multivariate components affect the distribution of the maximum statistic, and whether dimensionality interacts with the estimated roughness (using residual-based estimation [29]. In particular, if increasing dimensionality increases the kurtosis of the maximum-distribution more severely than the per-voxel distributions, then this could explain the relatively worse FWE performance.

One might argue that uncorrected p-values should be sufficient for methodological comparisons, since controlling the false-positive rate is not of particular importance when the results are only used to compare methods, and not reported as clinical findings. However, if different methods lead to different levels of spatial regularity in their results, this is of methodological interest. As argued by Poldrack et al. [121],[39] uncorrected p-values lead

[38] The distinction between 'precise' and 'accurate' registration must be emphasised here; we use the term precise with regard to fine-scale alignment and a sharp average atlas, but this does not imply accurate registration in terms of identifiable landmarks or true underlying anatomical correspondences.

[39] One of the coauthors, Brett, argues still more forcefully in unpublished comments online `http://imaging.mrc-cbu.cam.ac.uk/imaging/UncorrectedThreshold`.

to unquantified false-positive rates, due to the potential for different studies to have different effective degrees of multiplicity. As discussed in section 4.3.3 the signal-adaptivity of FDR procedures makes them problematic for method comparison. We therefore place more weight on our FWE-corrected results, which favour a balance between more information in the higher dimensional measures versus more reliable estimation and testing in the more parsimonious ones. There is a clear need for deeper understanding of the discrepancy though, so our results should not be overstated. At this stage, it seems fair to conclude only that Lepore et al. [23] might possibly have exaggerated the benefits of generalised TBM versus conventional scalar TBM, by not considering FWE-corrected significance.

As well as the many avenues for further experimental work and performance characterisation, there remain several open theoretical questions regarding the mathematics of Jacobian matrices. Some of this theory is expanded upon in chapter 5.

Bibliography

[1] N. Kriegeskorte, R. Goebel, and P. Bandettini, "Information-based functional brain mapping." *Proc Natl Acad Sci U S A*, vol. 103, no. 10, pp. 3863–3868, Mar. 2006. ˆ186, 187, 190, 191, 225, 288

[2] J. Ashburner, "Computational neuroanatomy," Ph.D. dissertation, University College London, 2000. [Online]. Available: http://www.fil.ion.ucl.ac.uk/spm/doc/theses/john/ ˆ186, 190, 191, 208, 230, 232

[3] C. Davatzikos, "Why voxel-based morphometric analysis should be used with great caution when characterizing group differences." *Neuroimage*, vol. 23, no. 1, pp. 17–20, Sep. 2004. ˆ186

[4] A. Mechelli, C. J. Price, K. J. Friston, and J. Ashburner, "Voxel-based morphometry of the human brain: Methods and applications," *Current Medical Imaging Reviews*, vol. 1, no. 1, pp. 1–9, 2005. ˆ186

[5] C. F. Beckmann and S. M. Smith, "Probabilistic independent component analysis for functional magnetic resonance imaging," *IEEE Trans. Med. Imag.*, vol. 23, no. 2, pp. 137–152, 2004. ˆ187

[6] ——, "Tensorial extensions of independent component analysis for multisubject FMRI analysis." *Neuroimage*, vol. 25, no. 1, pp. 294–311, Mar. 2005. ˆ187

[7] K. Friston, C. Chu, J. Mourão-Miranda, O. Hulme, G. Rees, W. Penny, and J. Ashburner, "Bayesian decoding of brain images." *Neuroimage*, vol. 39, no. 1, pp. 181–205, Jan. 2008. ˆ187

[8] J. M. Schott, S. L. Price, C. Frost, J. L. Whitwell, M. N. Rossor, and N. C. Fox, "Measuring atrophy in Alzheimer disease: a serial MRI study over 6 and 12 months." *Neurology*, vol. 65, no. 1, pp. 119–124, Jul. 2005. ˆ187, 223, 224

[9] H. J. Keselman, J. Algina, and R. K. Kowalchuk, "The analysis of repeated measures designs: a review." *Br J Math Stat Psychol*, vol. 54, no. Pt 1, pp. 1–20, May 2001. ˆ187

[10] S. Hayasaka, A.-T. Du, A. Duarte, J. Kornak, G.-H. Jahng, M. W. Weiner, and N. Schuff, "A non-parametric approach for co-analysis of multi-modal brain imaging data: application to Alzheimer's disease." *Neuroimage*, vol. 30, no. 3, pp. 768–779, Apr. 2006. ˆ188

[11] S. C. L. Deoni, T. M. Peters, and B. K. Rutt, "High-resolution T1 and T2 mapping of the brain in a clinically acceptable time with DESPOT1 and DESPOT2." *Magn Reson Med*, vol. 53, no. 1, pp. 237–241, Jan. 2005. ˆ188

[12] W. E. Klunk, H. Engler, A. Nordberg, Y. Wang, G. Blomqvist, D. P. Holt, M. Bergström, I. Savitcheva, G. feng Huang, S. Estrada, B. Ausén, M. L. Debnath, J. Barletta, J. C. Price, J. Sandell, B. J. Lopresti, A. Wall, P. Koivisto, G. Antoni, C. A. Mathis, and B. Långström, "Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B." *Ann Neurol*, vol. 55, no. 3, pp. 306–319, Mar. 2004. ˆ188

[13] J. Ashburner, C. Hutton, R. Frackowiak, I. Johnsrude, C. Price, and K. Friston, "Identifying global anatomical differences: deformation-based morphometry." *Hum Brain Mapp*, vol. 6, no. 5-6, pp. 348–357, 1998. [Online]. Available: http://www3.interscience.wiley.com/journal/79011/abstract ˆ189

[14] C. Gaser, H. P. Volz, S. Kiebel, S. Riehemann, and H. Sauer, "Detecting structural changes in whole brain based on nonlinear deformations-application to schizophrenia research." *Neuroimage*, vol. 10, no. 2, pp. 107–113, Aug. 1999. ˆ189, 190

[15] V. A. Cardenas, C. Studholme, S. Gazdzinski, T. C. Durazzo, and D. J. Meyerhoff, "Deformation-based morphometry of brain changes in alcohol dependence and abstinence." *Neuroimage*, vol. 34, no. 3, pp. 879–887, Feb. 2007. ˆ189

[16] B. Whitcher, J. J. Wisco, N. Hadjikhani, and D. S. Tuch, "Statistical group comparison of diffusion tensors via multivariate hypothesis testing." *Magn Reson Med*, vol. 57, no. 6, pp. 1065–1074, Jun. 2007. ˆ189, 208, 226, 232, 233, 262, 289

[17] K. Mardia, J. Kent, and J. Bibby, *Multivariate analysis*. Academic Press, 1979. ˆ189

[18] F. Pesarin, *Multivariate Permutation Tests: With Applications in Biostatistics*. J. Wiley, 2001. ˆ189

[19] A. D. Leow, A. D. Klunder, C. R. Jack, A. W. Toga, A. M. Dale, M. A. Bernstein, P. J. Britson, J. L. Gunter, C. P. Ward, J. L. Whitwell, B. J. Borowski, A. S. Fleisher, N. C. Fox, D. Harvey, J. Kornak, N. Schuff, C. Studholme, G. E. Alexander, M. W. Weiner, and P. M. Thompson, "Longitudinal stability of MRI for mapping brain

change using tensor-based morphometry." *Neuroimage*, vol. 31, no. 2, pp. 627–640, Jun. 2006, A.D.N.I. Preparatory Phase Study. ˆ189, 232

[20] C. Studholme, V. Cardenas, R. Blumenfeld, N. Schuff, H. J. Rosen, B. Miller, and M. Weiner, "Deformation tensor morphometry of semantic dementia with quantitative validation." *Neuroimage*, vol. 21, no. 4, pp. 1387–1398, Apr. 2004. ˆ189

[21] J. Ashburner and K. J. Friston, *Morphometry*, 2nd ed. Academic Press, 2004, ch. 6. [Online]. Available: http://www.fil.ion.ucl.ac.uk/spm/doc/books/hbf2/pdfs/Ch6.pdf ˆ189, 197, 198

[22] N. Lepore, C. A. Brun, M.-C. Chiang, Y.-Y. Chou, R. A. Dutton, K. M. Hayashi, O. L. Lopez, H. J. Aizenstein, A. W. Toga, J. T. Becker, and P. M. Thompson, "Multivariate statistics of the Jacobian matrices in Tensor Based Morphometry and their application to HIV/AIDS," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, ser. Springer LNCS, 2006, pp. 191–198. ˆ189, 191, 232, 249

[23] N. Lepore, C. Brun, Y. Y. Chou, M. C. Chiang, R. A. Dutton, K. M. Hayashi, E. Luders, O. L. Lopez, H. J. Aizenstein, A. W. Toga, J. T. Becker, and P. M. Thompson, "Generalized tensor-based morphometry of HIV/AIDS using multivariate statistics on deformation tensors." *IEEE Trans. Med. Imag.*, vol. 27, no. 1, pp. 129–141, Jan. 2008. ˆ189, 191, 197, 200, 208, 209, 211, 212, 229, 230, 232, 249, 286, 291

[24] C. Studholme and V. Cardenas, "Population based analysis of directional information in serial deformation tensor morphometry." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 10, no. Pt 2, 2007, pp. 311–318. [Online]. Available: http://www.springerlink.com/content/p7v37313067q2472/ ˆ190, 191, 203, 216, 217

[25] N. Kriegeskorte and P. Bandettini, "Analyzing for information, not activation, to exploit high-resolution fMRI." *Neuroimage*, vol. 38, no. 4, pp. 649–662, Dec. 2007. ˆ190, 191

[26] C. Gaser, I. Nenadic, B. R. Buchsbaum, E. A. Hazlett, and M. S. Buchsbaum, "Deformation-based morphometry and its relation to conventional volumetry of brain lateral ventricles in MRI." *Neuroimage*, vol. 13, no. 6 Pt 1, pp. 1140–1145, Jun. 2001. ˆ190, 232

[27] J. Cao and K. Worsley, "The detection of local shape changes via the geometry of Hotelling's T^2 fields," *The Annals of Statistics*, vol. 27, no. 3, pp. 925–942, 1999. ˆ190

[28] F. Carbonell, L. Galan, and K. Worsley, "The geometry of the Wilks's lambda random field," *Annals of the institute of Statistical Mathematics*, 2007, accepted. [Online]. Available: http://www.math.mcgill.ca/keith/felix/felix.htm ˆ190

[29] K. Worsley, "An unbiased estimator for the roughness of a multivariate Gaussian random field," Department of Mathematics and Statistics, McGill University, Canada, Tech. Rep., 1996. [Online]. Available: http://www.math.mcgill.ca/~keith/smoothness/techrept.pdf ˆ190, 230, 257, 260, 290

[30] S. J. Kiebel, J. B. Poline, K. J. Friston, A. P. Holmes, and K. J. Worsley, "Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model." *Neuroimage*, vol. 10, no. 6, pp. 756–766, Dec. 1999. ˆ190

[31] J. D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 479–498, 2002. [Online]. Available: http://www.jstor.org/stable/3088784 ˆ191

[32] ——, "The positive false discovery rate: A bayesian interpretation and the q-value," *The Annals of Statistics*, vol. 31, no. 6, pp. 2013–2035, Dec. 2003. [Online]. Available: http://www.jstor.org/stable/3448445 ˆ191

[33] C. R. Genovese, N. A. Lazar, and T. Nichols, "Thresholding of statistical maps in functional neuroimaging using the false discovery rate." *Neuroimage*, vol. 15, no. 4, pp. 870–878, Apr. 2002. ˆ191, 229, 230

[34] K. Worsley, S. Marrett, P. Neelin, and A. Evans, "Searching scale space for activation in PET images," *Human Brain Mapping*, vol. 4, no. 1, pp. 74–90, 1996. ˆ192, 235

[35] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA Engineer*, vol. 29, no. 6, 1984. [Online]. Available: http://citeseer.ist.psu.edu/adelson84pyramid.html ˆ192

[36] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images." *Neuroimage*, vol. 17, no. 2, pp. 825–841, Oct. 2002. ˆ192

[37] J. Ashburner, "A fast diffeomorphic image registration algorithm." *Neuroimage*, vol. 38, no. 1, pp. 95–113, Oct. 2007. ˆ192, 221, 222, 287, 290

[38] P. Thevenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Trans. Image Process.*, vol. 9, no. 12, pp. 2083–2099, 2000. ˆ192

[39] P. Vemuri, J. L. Gunter, M. L. Senjem, J. L. Whitwell, K. Kantarci, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack, "Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies." *Neuroimage*, vol. 39, no. 3, pp. 1186–1197, Feb. 2008. ˆ192

[40] Z. Lao, D. Shen, Z. Xue, B. Karacali, S. M. Resnick, and C. Davatzikos, "Morphological classification of brains via high-dimensional shape transformations and machine learning methods." *Neuroimage*, vol. 21, no. 1, pp. 46–57, Jan. 2004. ˆ192

[41] Y. Fan, D. Shen, R. C. Gur, R. E. Gur, and C. Davatzikos, "COMPARE: Classification of morphological patterns using adaptive regional elements," *IEEE Trans. Med. Imag.*, vol. 26, no. 1, pp. 93–105, 2007. ˆ192

[42] M. Unser, A. Aldroubi, and M. Eden, "The l2-polynomial spline pyramid," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 4, pp. 364–379, 1993. ˆ192, 247

[43] P. Brigger, F. Muller, K. Illgner, and M. Unser, "Centered pyramids," *IEEE Trans. Image Process.*, vol. 8, no. 9, pp. 1254–1264, 1999. ˆ192, 247

[44] E. Bullmore, C. Long, J. Suckling, J. Fadili, G. Calvert, F. Zelaya, T. A. Carpenter, and M. Brammer, "Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains." *Hum Brain Mapp*, vol. 12, no. 2, pp. 61–78, Feb. 2001. ˆ192

[45] E. Bullmore, J. Fadili, V. Maxim, L. Sendur, B. Whitcher, J. Suckling, M. Brammer, and M. Breakspear, "Wavelets and functional magnetic resonance imaging of the human brain." *Neuroimage*, vol. 23 Suppl 1, pp. S234–S249, 2004. ˆ192

[46] A. M. Wink and J. B. T. M. Roerdink, "Denoising functional MR images: a comparison of wavelet denoising and Gaussian smoothing," *IEEE Trans. Med. Imag.*, vol. 23, no. 3, pp. 374–387, 2004. ˆ192

[47] A. F. Bower, "Applied mechanics of solids," 2008, online Textbook. [Online]. Available: http://solidmechanics.org/ ˆ194, 197, 198, 199, 209

[48] J. Ashburner and K. J. Friston, "Computing average shaped tissue probability templates." *Neuroimage*, vol. 45, no. 2, pp. 333–341, Apr. 2009. ˆ194, 290

[49] R. G. Boyes, D. Rueckert, P. Aljabar, J. Whitwell, J. M. Schott, D. L. G. Hill, and N. C. Fox, "Cerebral atrophy measurements using Jacobian integration: comparison with the boundary shift integral." *Neuroimage*, vol. 32, no. 1, pp. 159–169, Aug. 2006. ˆ194

[50] M. K. Chung, K. J. Worsley, T. Paus, C. Cherif, D. L. Collins, J. N. Giedd, J. L. Rapoport, and A. C. Evans, "A unified statistical approach to deformation-based morphometry." *Neuroimage*, vol. 14, no. 3, pp. 595–606, Sep. 2001. ˆ194, 196, 199, 209, 217, 280, 281

[51] K. Riley, M. Hobson, and S. Bence, *Mathematical Methods for Physics and Engineering*, 3rd ed. Cambridge University Press, 2006. ˆ195

[52] X. Pennec, R. Stefanescu, V. Arsigny, P. Fillard, and N. Ayache, "Riemannian elasticity: a statistical regularization framework for non-linear registration." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 8, no. Pt 2, 2005, pp. 943–950. [Online]. Available: http://www.springerlink.com/content/2vyqc0lqgwkh3e2k/ ˆ197, 198

[53] B. C. Wünsche, R. Lobb, and A. A. Young, "The visualization of myocardial strain for the improved analysis of cardiac mechanics," in *2nd international conference on Computer graphics and interactive techniques*. ACM, 2004, pp. 90–99. ˆ197, 287

[54] C. Le Guyader and L. Vese, "A combined segmentation and registration framework with a nonlinear elasticity smoother," University of California, Los Angeles, Tech. Rep. 08-16, Mar. 2008. [Online]. Available: http://www.math.ucla.edu/applied/cam ˆ198

[55] J. Ashburner, J. L. Andersson, and K. J. Friston, "Image registration using a symmetric prior–in three dimensions." *Hum Brain Mapp*, vol. 9, no. 4, pp. 212–225, Apr. 2000. [Online]. Available: http://www3.interscience.wiley.com/journal/71001030/abstract ˆ199, 273

[56] R. P. Woods, "Characterizing volume and surface deformations in an atlas framework: theory, applications, and implementation." *Neuroimage*, vol. 18, no. 3, pp. 769–788, Mar. 2003. ˆ200, 202, 206, 208

[57] M. Moakher, "Means and averaging in the group of rotations," *SIAM Journal on matrix analysis and applications*, vol. 24, no. 1, pp. 1–16, 2002. ˆ201

[58] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian framework for tensor computing," *International Journal of Computer Vision*, vol. 66, no. 1, pp. 41–66, 2006. ˆ203, 205, 206, 232, 289

[59] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 29, no. 1, p. 328, 2008. ˆ203, 206, 207, 232

[60] P. G. Batchelor, M. Moakher, D. Atkinson, F. Calamante, and A. Connelly, "A rigorous framework for diffusion tensor calculus." *Magn Reson Med*, vol. 53, no. 1, pp. 221–225, Jan. 2005. ˆ205, 206, 210, 211, 212, 289

[61] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Log-Euclidean metrics for fast and simple calculus on diffusion tensors." *Magn Reson Med*, vol. 56, no. 2, pp. 411–421, Aug. 2006. ˆ207

[62] P. J. Basser and C. Pierpaoli, "Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI." *J Magn Reson B*, vol. 111, no. 3, pp. 209–219, Jun. 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8661285 ˆ210, 211, 212, 288

[63] D. Alexander, C. Pierpaoli, P. Basser, and J. Gee, "Spatial transformations of diffusion tensor magnetic resonance images," *IEEE Trans. Med. Imag.*, vol. 20, no. 11, pp. 1131–1139, 2001. ˆ210, 216, 217, 218, 220, 221

[64] A. Schwartzman, R. F. Dougherty, and J. E. Taylor, "Cross-subject comparison of principal diffusion direction maps." *Magn Reson Med*, vol. 53, no. 6, pp. 1423–1431, Jun. 2005. ˆ213, 226, 227, 230, 289

[65] C. Frost, M. G. Kenward, and N. C. Fox, "The analysis of repeated 'direct' measures of change illustrated with an application in longitudinal imaging." *Stat Med*, vol. 23, no. 21, pp. 3275–3286, Nov. 2004. ˆ214

[66] P. A. Freeborough and N. C. Fox, "The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI." *IEEE Trans. Med. Imag.*, vol. 16, no. 5, pp. 623–629, Oct. 1997. ˆ214

[67] A. Rao, R. Chandrashekara, G. Sanchez-Ortiz, R. Mohiaddin, P. Aljabar, J. Hajnal, B. Puri, and D. Rueckert, "Spatial transformation of motion and deformation fields using nonrigid registration," *IEEE Trans. Med. Imag.*, vol. 23, no. 9, pp. 1065–1076, Sep. 2004. ˆ214, 215, 216, 217, 218, 220, 222

[68] D. Xu, S. Mori, D. Shen, P. C. M. van Zijl, and C. Davatzikos, "Spatial normalization of diffusion tensor fields." *Magn Reson Med*, vol. 50, no. 1, pp. 175–182, Jul. 2003. ˆ219

[69] D. Xu, X. Hao, R. Bansal, K. J. Plessen, and B. S. Peterson, "Seamless warping of diffusion tensor fields." *IEEE Trans. Med. Imag.*, vol. 27, no. 3, pp. 285–299, Mar. 2008. ˆ219, 221, 222

[70] H. Zhang, P. A. Yushkevich, D. C. Alexander, and J. C. Gee, "Deformable registration of diffusion tensor MR images with explicit orientation optimization." *Med Image Anal*, vol. 10, no. 5, pp. 764–785, Oct. 2006. ˆ220

[71] W. R. Crum, O. Camara, and D. J. Hawkes, "Methods for inverting dense displacement fields: evaluation in brain image registration." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 10, no. Pt 1, 2007, pp. 900–907. [Online]. Available: http://www.springerlink.com/content/132m22p1w68p1350/ ˆ222

[72] D. Rueckert, P. Aljabar, R. A. Heckemann, J. V. Hajnal, and A. Hammers, "Diffeomorphic registration using B-splines." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 9, no. Pt 2, 2006, pp. 702–709. [Online]. Available: http://www.springerlink.com/content/u11268n6734lpg41/ ˆ222

[73] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Non-parametric diffeomorphic image registration with the demons algorithm." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 10, no. Pt 2, 2007, pp. 319–326. [Online]. Available: http://www.springerlink.com/content/0uj2712ju7r554q1/ ˆ222

[74] J. Ashburner, "Using DARTEL," 2008, software documentation. [Online]. Available: http://www.fil.ion.ucl.ac.uk/~john/misc/dartel_guide.pdf ˆ222

[75] J. Ashburner and K. J. Friston, "Unified segmentation." *Neuroimage*, vol. 26, no. 3, pp. 839–851, Jul. 2005. ˆ223

[76] P. Thevenaz, T. Blu, and M. Unser, "Interpolation revisited," *IEEE Trans. Med. Imag.*, vol. 19, no. 7, pp. 739–758, 2000. ˆ223

[77] J. Ashburner, J. L. Andersson, and K. J. Friston, "High-dimensional image registration using symmetric priors." *Neuroimage*, vol. 9, no. 6 Pt 1, pp. 619–628, Jun. 1999. ˆ223

[78] J. Barnes, S. Henley, M. Lehmann, N. Hobbs, R. Scahill, M. Clarkson, G. Ridgway, D. MacManus, S. Ourselin, and N. Fox, "Head size, age and gender adjustment in MRI studies: A necessary nuisance?" *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, vol. 5, no. 4S, pp. 102–103, 2009. ˆ224

[79] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, 2004. ˆ225

[80] C. Thomaz, D. Gillies, and R. Feitosa, "A new covariance estimate for bayesian classifiers in biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 2, pp. 214–223, 2004. ˆ225

[81] C. E. Thomaz, J. P. Boardman, D. L. G. Hill, J. V. Hajnal, A. D. Edwards, M. A. Rutherford, D. F. Gillies, and D. Rueckert, "Using a maximum uncertainty LDA-based approach to classify and analyse MR brain images," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 3216. LNCS, 2004, pp. 291–300. ˆ225

[82] L. Baringhaus and C. Franz, "On a new multivariate two-sample test," *Journal of Multivariate Analysis*, vol. 88, no. 1, pp. 190–206, 2004. ˆ225, 226

[83] S. LaConte, J. Anderson, S. Muley, J. Ashe, S. Frutiger, K. Rehm, L. K. Hansen, E. Yacoub, X. Hu, D. Rottenberg, and S. Strother, "The evaluation of preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics." *Neuroimage*, vol. 18, no. 1, pp. 10–27, Jan. 2003. ˆ228

[84] S. Strother, "Evaluating fMRI preprocessing pipelines," *IEEE Eng. Med. Biol. Mag.*, vol. 25, no. 2, pp. 27–41, 2006. ˆ228

[85] X. Hua, A. D. Leow, S. Lee, A. D. Klunder, A. W. Toga, N. Lepore, Y.-Y. Chou, C. Brun, M.-C. Chiang, M. Barysheva, C. R. Jack, M. A. Bernstein, P. J. Britson, C. P. Ward, J. L. Whitwell, B. Borowski, A. S. Fleisher, N. C. Fox, R. G. Boyes, J. Barnes, D. Harvey, J. Kornak, N. Schuff, L. Boreta, G. E. Alexander, M. W. Weiner, P. M. Thompson, and the Alzheimer's Disease Neuroimaging Initiative, "3D characterization of brain atrophy in Alzheimer's disease and mild cognitive impairment using tensor-based morphometry." *Neuroimage*, vol. 41, no. 1, pp. 19–34, May 2008. ˆ228, 229

[86] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement." *Lancet*, vol. 1, no. 8476, pp. 307–310, Feb. 1986. ˆ229

[87] T. Nichols and S. Hayasaka, "Controlling the familywise error rate in functional neuroimaging: a comparative review." *Stat Methods Med Res*, vol. 12, no. 5, pp. 419–446, Oct. 2003. ˆ230

[88] P. Cachier and D. Rey, "Symmetrization of the non-rigid registration problem using inversion-invariant energies: Application to multiple sclerosis," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, ser. LNCS. Springer, 2000, pp. 472–481. [Online]. Available: http://www.springerlink.com/content/0b4dn00pucxunpct/ ˆ232

[89] T. Rohlfing, E. V. Sullivan, and A. Pfefferbaum, "Deformation-based brain morphometry to track the course of alcoholism: differences between intra-subject and inter-subject analysis." *Psychiatry Res*, vol. 146, no. 2, pp. 157–170, Mar. 2006. ˆ232

[90] A. D. Leow, I. Yanovsky, M.-C. Chiang, A. D. Lee, A. D. Klunder, A. Lu, J. T. Becker, S. W. Davis, A. W. Toga, and P. M. Thompson, P. M. A10 Thompson, "Statistical properties of Jacobian maps and the realization of unbiased large-deformation nonlinear image registration," *IEEE Trans. Med. Imag.*, vol. 26, no. 6, pp. 822–832, 2007. ˆ232

[91] M. Moakher, "A differential geometric approach to the geometric mean of symmetric positive-definite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 26, no. 3, p. 735, 2005. ˆ232

[92] M. Belmonte and D. Yurgelun-Todd, "Permutation testing made practical for functional magnetic resonance image analysis," *IEEE Trans. Med. Imag.*, vol. 20, no. 3, pp. 243–248, Mar. 2001. ˆ234, 269

[93] T. E. Nichols and A. P. Holmes, "Nonparametric permutation tests for functional neuroimaging: A primer with examples," *Human Brain Mapping*, vol. 15, no. 1, pp. 1–25, 2002. ˆ234, 287

[94] R. I. Scahill, "MRI in Alzheimer's disease and related disorders : the application of statistical tests of difference to serial and cross-sectional imaging to improve diagnosis and progression measurement," Ph.D. dissertation, University College London, 2003. ˆ235

[95] D. K. Jones, M. R. Symms, M. Cercignani, and R. J. Howard, "The effect of filter size on VBM analyses of DT-MRI data." *Neuroimage*, vol. 26, no. 2, pp. 546–554, Jun. 2005. ˆ235, 264

[96] M. Reimold, M. Slifstein, A. Heinz, W. Mueller-Schauenburg, and R. Bares, "Effect of spatial smoothing on t-maps: arguments for going back from t-maps to masked

contrast images." *J Cereb Blood Flow Metab*, vol. 26, no. 6, pp. 751–759, Jun. 2006. ˆ237

[97] S. D. Forman, J. D. Cohen, M. Fitzgerald, W. F. Eddy, M. A. Mintun, and D. C. Noll, "Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold." *Magn Reson Med*, vol. 33, no. 5, pp. 636–647, May 1995. ˆ241

[98] C. Poupon, C. A. Clark, V. Frouin, J. Rgis, I. Bloch, D. L. Bihan, and J. Mangin, "Regularization of diffusion-based direction maps for the tracking of brain white matter fascicles." *Neuroimage*, vol. 12, no. 2, pp. 184–195, Aug. 2000. ˆ275

[99] M. Hernandez, M. N. Bossa, and S. Olmos, "Registration of anatomical images using geodesic paths of diffeomorphisms parameterized with stationary vector fields," in *Proc. IEEE 11th International Conference on Computer Vision*, M. N. Bossa, Ed., 2007, pp. 1–8. ˆ287

[100] V. Arsigny, O. Commowick, X. Pennec, and N. Ayache, "A log-Euclidean framework for statistics on diffeomorphisms." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 9, no. Pt 1, 2006, pp. 924–931. [Online]. Available: http://www.springerlink.com/content/607206763v078397/ ˆ287

[101] M. Beg, M. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *International Journal of Computer Vision*, vol. 61, no. 2, pp. 139–157, 2005. ˆ287

[102] M. Miller, A. Trouvé, and L. Younes, "Geodesic shooting for computational anatomy," *Journal of Mathematical Imaging and Vision*, vol. 24, no. 2, pp. 209–228, 2006. [Online]. Available: http://www.springerlink.com/content/9r82230441886375/ ˆ287

[103] S. Marsland and R. McLachlan, "A Hamiltonian particle method for diffeomorphic image registration." in *Inf. Process. Med. Imag.*, vol. 20, 2007, pp. 396–407. [Online]. Available: http://www.springerlink.com/content/x5q066610172r113/ ˆ287

[104] L. Wang, M. Beg, J. Ratnanather, C. Ceritoglu, L. Younes, J. C. Morris, J. G. Csernansky, and M. I. Miller, "Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type," *IEEE Trans. Med. Imag.*, vol. 26, no. 4, p. 462, 2007. ˆ287

[105] E. Bullmore, J. Suckling, S. Overmeyer, S. Rabe-Hesketh, E. Taylor, and M. Brammer, "Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain," *IEEE Trans. Med. Imag.*, vol. 18, no. 1, pp. 32–42, Jan. 1999. ˆ287

[106] J. Suckling and E. Bullmore, "Permutation tests for factorially designed neuroimaging experiments." *Hum Brain Mapp*, vol. 22, no. 3, pp. 193–205, Jul. 2004. ˆ287

[107] S. Hayasaka and T. E. Nichols, "Combining voxel intensity and cluster extent with permutation test framework." *Neuroimage*, vol. 23, no. 1, pp. 54–63, Sep. 2004. ˆ288

[108] K. J. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner, "Classical and Bayesian inference in neuroimaging: theory." *Neuroimage*, vol. 16, no. 2, pp. 465–483, Jun. 2002. ˆ288

[109] M. W. Woolrich, M. Jenkinson, J. M. Brady, and S. M. Smith, "Fully Bayesian spatio-temporal modeling of FMRI data." *IEEE Trans. Med. Imag.*, vol. 23, no. 2, pp. 213–231, Feb. 2004. ˆ288

[110] M. W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, and S. M. Smith, "Bayesian analysis of neuroimaging data in FSL," *Neuroimage*, Nov. 2008. ˆ288

[111] W. D. Penny, N. J. Trujillo-Barreto, and K. J. Friston, "Bayesian fMRI time series analysis with spatial priors." *Neuroimage*, vol. 24, no. 2, pp. 350–362, Jan. 2005. ˆ288

[112] M. Woolrich, T. Behrens, C. Beckmann, and S. Smith, "Mixture models with adaptive spatial regularization for segmentation with an application to FMRI data," *IEEE Trans. Med. Imag.*, vol. 24, no. 1, pp. 1–11, Jan. 2005. ˆ288

[113] W. Penny, G. Flandin, and N. Trujillo-Barreto, "Bayesian comparison of spatially regularised general linear models." *Hum Brain Mapp*, vol. 28, no. 4, pp. 275–293, Apr. 2007. ˆ288

[114] G. Flandin and W. D. Penny, "Bayesian fMRI data analysis with sparse spatial basis function priors." *Neuroimage*, vol. 34, no. 3, pp. 1108–1125, Feb. 2007. ˆ288

[115] S. Makni, C. Beckmann, S. Smith, and M. Woolrich, "Bayesian deconvolution fMRI data using bilinear dynamical systems." *Neuroimage*, vol. 42, no. 4, pp. 1381–1396, Oct. 2008. ˆ288

[116] L. M. Harrison, W. Penny, J. Ashburner, N. Trujillo-Barreto, and K. J. Friston, "Diffusion-based spatial priors for imaging." *Neuroimage*, vol. 38, no. 4, pp. 677–695, Dec. 2007. ˆ288

[117] L. M. Harrison, W. Penny, J. Daunizeau, and K. J. Friston, "Diffusion-based spatial priors for functional magnetic resonance images." *Neuroimage*, vol. 41, no. 2, pp. 408–423, Jun. 2008. ˆ288

[118] L. M. Harrison, W. Penny, G. Flandin, C. C. Ruff, N. Weiskopf, and K. J. Friston, "Graph-partitioned spatial priors for functional magnetic resonance images." *Neuroimage*, vol. 43, no. 4, pp. 694–707, Dec. 2008. ˆ288

[119] G. Gerig, O. Kubler, R. Kikinis, and F. Jolesz, "Nonlinear anisotropic filtering of MRI data," *IEEE Trans. Med. Imag.*, vol. 11, no. 2, pp. 221–232, Jun. 1992. ˆ288

[120] S. Joshi, B. Davis, M. Jomier, and G. Gerig, "Unbiased diffeomorphic atlas construction for computational anatomy." *Neuroimage*, vol. 23 Suppl 1, pp. S151–S160, 2004. ˆ290

[121] R. A. Poldrack, P. C. Fletcher, R. N. Henson, K. J. Worsley, M. Brett, and T. E. Nichols, "Guidelines for reporting an fMRI study." *Neuroimage*, Dec. 2007. ˆ290

Chapter 5

Further Developments

This final chapter presents two distinct contributions which are in fact only the beginnings of work for which much more is planned. Neither section is complete, but they are included in the hope that they complement the more thorough investigations presented earlier in the thesis.

Firstly, the interesting mathematical issues surrounding analysis of the full Jacobian tensor are returned to, in a section which attempts to include both a review of some of the more theoretical issues, and some more intuitive discussion and practical examples that should help to put the theory in context.

Secondly, we thoroughly review the literature and begin to propose new methods within the important topic of differential bias correction. Longitudinal registration is the essential foundation of both chapters 3 and 4 (as well as much more beyond the scope of the present thesis), and the incorporation of better modelling or correction of differential bias should provide substantial improvements to this key technology.

5.1 Further tensor-based morphometry theory

Chapter 4 provided a reasonably thorough presentation of the theory for multivariate tensor-based morphometry, including some of the more mathematical aspects related to Riemannian geometry and suitable distance metrics. However, there are many issues that were not adequately analysed, and a number of possible practical methods for TBM analysis that we did not have time for in section 4.3. We now attempt to probe deeper into some of these mathematical issues, without performing any further experimental TBM work, with the main purpose of clarifying possible paths for future developments.

5.1.1 Distances and means for Jacobian matrices

In section 4.2.6, it was mentioned that there is no bi-invariant Riemannian metric for general matrices with positive determinant, such as Jacobian matrices [1]. More formally, we consider the group of 3×3 real matrices with positive determinant, denoted $GL^+(3, \mathbb{R})$ or $GL^+(3)$ for brevity. We will now discuss this in greater detail. Consider a cross-sectional data-set, where registrations have been computed from a chosen target to each of a set of n

images. Denoting the target as image 1 (so T_{11} is an identity transformation), this gives rise to a set of Jacobian matrices at each voxel $\{J_{11}(x,y,z) = I, J_{21}(x,y,z), \ldots, J_{n1}(x,y,z)\}$.[1] Since the choice of target image is arbitrary, it is desirable that an analysis based on these Jacobian matrices should be invariant to this choice. If one considers changing to the second image as the target, then the new transformations can be computed by composing the transformations from image 1 to each other image with the transformation from image 2 to image 1: $T_{i2} = T_{i1}T_{12} = T_{i1}T_{21}^{-1}$, which results in the new set of Jacobian matrices becoming $\{J_{12} = J_{21}^{-1}, J_{22} = I, J_{32} = J_{31}J_{21}^{-1} \ldots, J_{n2} = J_{n1}J_{21}^{-1}\}$, i.e. each matrix has been right-multiplied by J_{21}^{-1}. If the distance between Jacobians can be quantified with a metric that is invariant to such right-multiplication (i.e. a right-invariant metric) then the resulting Fréchet mean is also right-invariant, in the sense that the mean of $\{J_{ij}Q\}_{i=1}^{n}$ will be given by the mean of $\{J_{ij}\}_{i=1}^{n}$ right-multiplied by Q. The Jacobian matrices corresponding to the transformations from a hypothetical mean image to each of the images are then given by $J_{im} = J_{ij}J_{mj}^{-1}$, which are invariant to the right-multiplication (of both J_{ij} and J_{mj}) by Q, or equivalently, invariant to the choice of image j.

In addition to the arbitrary choice of target image, one could argue that the decision to analyse transformations from target to each other image, instead of from each image to the target, is also arbitrary. The set of Jacobians that arise from considering the mapping in the other direction are inverted. E.g. using image 1 as the target again yields: $\{J_{11}^{-1} = I, J_{12} = J_{21}^{-1}, \ldots, J_{1n} = J_{n1}^{-1}\}$. Woods [1] argues that the mean of these inverted matrices should equal the inverse of the mean of the originals, and hence that the distance metric should be invariant under inversion; Arsigny proves that right-invariance and inversion-invariance together imply bi-invariance (i.e. left- and right-invariance) [2].

Woods [1] states that there is no bi-invariant Riemannian metric for matrices in $GL^{+}(3)$, and Arsigny [2] proves that no metric exists for the special case of rigid body motions in any dimension. Taking the invariance of the mean as being more important than the metric property of the distance, Woods relaxes the requirement for a metric on a Riemannian manifold to allow a pseudo-metric on a semi-Riemannian manifold [1]. The pseudo-metric can be negative, meaning that distinct points can be separated by zero or negative 'distance' — i.e. the concept of distance has been lost. The concept of the Fréchet mean as the point which minimises the squared distances from itself to the observations is not applicable in the semi-Riemannian case. However, it is still possible to define geodesics as paths of constant velocity.[2] In Riemannian manifolds (/Euclidean space) the Fréchet (/arithmetic) mean also has the property that vectors in the tangent space (/original space) from the mean to each observation sum to zero. The constant velocity property of geodesics in semi-Riemannian manifolds allows the definition of a local 'Karcher mean' as a point from which velocity vectors in the tangent space associated with geodesics to each observation sum to zero. Woods furthermore states that the 'barycentric equation' (given

[1] We will henceforth drop the dependence on voxel (x, y, z).

[2] A simple analogy might help to clarify this rather abstract concept. Consider satellites orbiting the Earth, they trace great circles (like the equator) which are geodesics within the surface of the sphere. Not only are great circles distance-minimising geodesics, but they also have constant velocity in the tangent plane to the sphere, i.e. although the 3D motion of the satellite features centripetal acceleration towards the centre of the Earth, its acceleration in the two perpendicular dimensions is zero.

in section 4.2.7) can be used to find such a mean, provided the observations are 'close' enough on the manifold,[3] [1]

$$\bar{J}_{t+1} = \operatorname{expm}\left(\frac{1}{N}\sum_{i=1}^{N}\operatorname{logm}\left(J_i\bar{J}_t^{-1}\right)\right)\bar{J}_t.$$

Woods [1] shows that Jacobian matrices in $GL^+(3)$ can be embedded in the semi-simple Lie group $SL(4)$ (of 4×4 unity-determinant matrices) and states that semi-simple Lie groups have bi-invariant pseudo-metrics, suggesting (somewhat implicitly) that the barycentric equation therefore yields a bi-invariant mean. Arsigny's PhD thesis [2] contains an entire chapter on 'Bi-Invariant Means in Lie Groups', deriving them from a (Lie) algebraic perspective, rather than from the (semi-Riemannian) geometric perspective used by Woods.

Deviations from the bi-invariant mean

Woods framework initially seems theoretically very appealing. However, the mean Jacobian is itself of very limited practical interest. Morphometry requires more general analysis of the distribution of Jacobians, for example, the difference of the means of two or more groups, or the correlation of the distribution of Jacobians with some other covariates. Woods suggests that 'deviations' from the mean can be characterised with the 'initial velocities of the acceleration-free geodesics that carry the mean Jacobian matrix to the individual Jacobian matrices' [1]

$$X_i = \operatorname{logm}\left(J_i\bar{J}^{-1}\right). \tag{5.1}$$

These X_i lie in the tangent space, and can hence be treated as points in a nine-dimensional Euclidean vector space.

We observe two problems with this suggestion. Firstly, the use of multivariate statistics on these deviations seems to implicitly assume that their norm can be treated as a distance from the mean. For example, Woods suggests principal component analysis (sometimes known as principal geodesic analysis when performed on a manifold [3]) can be performed on the X_i; however, the principal directions would intuitively be those along which the distances $\|X_i\|$ would be maximised — yet these are not distances, since the (Frobenius) norm of (5.1) is not a metric, as shown in section 4.2.6. Arsigny's thesis provides further reason to doubt this approach, explaining that although bi-invariant means can be defined without a metric, higher order moments do require a Riemannian metric, because they involve inner products of vectors — a concept intimately connected with norms and hence metrics for differences in vectors [2]. Arsigny states that left- or right-invariant Riemannian metrics may be used where no bi-invariant one is available, but it seems that Woods' deviations are not based on such a metric.[4] A second apparent problem with

[3] 'Close' might sound a poorly defined concept given the lack of distance metric, however Woods provides a procedure for checking that the mean is unique, as well as rough arguments regarding its existence. A more thorough mathematical analysis of existence and uniqueness has been given by Arsigny [2].

[4] In fact, Woods does not explicitly state the form of the bi-invariant pseudo-metric on which the deviations are based either [1].

the deviations in (5.1) is that although they involve the bi-invariant mean, they do not appear to be bi-invariant themselves. This is surprising, given Woods emphasis on the need for bi-invariance, and his decision to abandon the concept of a metric in order to achieve it. Considering transforming the set of Jacobian matrices to $\{PJ_{ij}Q\}_{i=1}^{n}$, their mean transforms to $P\bar{J}Q$, giving

$$\begin{aligned} X_i^{PQ} &= \operatorname{logm}\left(PJ_i\,Q\left(P\bar{J}Q\right)^{-1}\right) \\ &= \operatorname{logm}\left(PJ_i\bar{J}^{-1}P^{-1}\right) \\ &= P\operatorname{logm}\left(J_i\bar{J}^{-1}\right)P^{-1}. \end{aligned}$$

Hence the deviations are right-invariant (since this corresponds to $P = I$), but generally not bi-invariant. Ignoring the complication that the norm of the deviation cannot be interpreted as a distance, we also observe that

$$\begin{aligned} \|X_i\|_F^2 &= \operatorname{tr}\left(\left(\operatorname{logm}\left(J_i\bar{J}^{-1}\right)\right)^T\operatorname{logm}\left(J_i\bar{J}^{-1}\right)\right); \\ \|X_i^{PQ}\|_F^2 &= \operatorname{tr}\left(\left(P\operatorname{logm}\left(J_i\bar{J}^{-1}\right)P^{-1}\right)^T\left(P\operatorname{logm}\left(J_i\bar{J}^{-1}\right)P^{-1}\right)\right) \\ &= \operatorname{tr}\left(P^{-T}\left(\operatorname{logm}\left(J_i\bar{J}^{-1}\right)\right)^T P^TP\operatorname{logm}\left(J_i\bar{J}^{-1}\right)P^{-1}\right) \\ &= \operatorname{tr}\left(\left(P^TP\right)^{-1}\left(\operatorname{logm}\left(J_i\bar{J}^{-1}\right)\right)^T\left(P^TP\right)\operatorname{logm}\left(J_i\bar{J}^{-1}\right)\right), \end{aligned}$$

which are only identical in general if $P^TP = sI$, i.e. the norm of the deviation is left-invariant to geometric similarity transformations.

Log-Euclidean analysis of Jacobian matrices

While Woods [1] opted for a semi-Riemannian analysis of Jacobian matrices, in section 4.2.7 we chose to focus on symmetric positive definite (SPD) strain tensors, for which a bi-invariant Riemannian metric is available. Furthermore, motivated by reduced computational burden, and the fact that bi-invariance is less important for longitudinal TBM, we elected to use the simpler log-Euclidean framework, as in [4].

Since we have shown in the preceding subsection that Woods' framework does not in fact obtain complete bi-invariance, the log-Euclidean approach deserves closer consideration. In particular, it is natural to ask whether the log-Euclidean metric is suitable for general Jacobian matrices in $GL^+(3)$ rather than SPD matrices. The answer is yes, but with some caveats, which we will now discuss. Firstly, we note that Lepore et al. [4] did discuss this possibility:

> By examining the deformation tensor in this work, we are examining only the SPD part of the Jacobian matrix, and three remaining degrees of freedom (a rotational term) are still discarded and not used. The Log-Euclidean framework can be extended to analyse the full Jacobian matrices, performing computations on that space (see [2] for extensions of the Log-Euclidean framework to general matrix spaces).

They go on to argue that the three additional parameters may increase power (as appears to be the case from our experimental results) but may also require a greater number of images.

Arsigny [2] investigates the log-Euclidean metric, and corresponding mean, for general linear transformation matrices, in the context of a poly-affine registration framework. Arsigny's group appear not to have applied their techniques to TBM, though they have investigated the morphometry of (manually-traced) sulci [5]. In addition to the reduced number of invariance properties, Arsigny notes that the log-Euclidean mean is limited to transformations close enough to the identity, while the bi-invariant mean of rigid transformations 'exists if and only if the bi-invariant mean of their rotation parts exists' [2]. In simple terms, we believe the bi-invariant mean requires $\operatorname{logm}\left(J_i\bar{J}^{-1}\right)$ is real for all J_i, while the log-Euclidean mean requires that $\operatorname{logm}\left(J_i\right)$ itself is real. For matrices (or quotients of matrices like $J_i\bar{J}^{-1}$) with positive determinants, eigenvalues must either be positive, paired complex conjugates, or repeated values on the negative-real line; only the final case precludes existence of a real matrix logarithm, and this case corresponds to rotations of 180° [1]. In other words, a unique bi-invariant mean exists so long as the rotational parts of the transformations are close enough to each other, while the log-Euclidean mean requires them also to be close enough to the identity. To give two simple examples, rotations around the z-axis of 170° and -170° have a bi-invariant mean of 180°, while rotations of 90° and -90° do not have a unique mean, since both 0° and 180° are at equal distance [6].

Intuitively, it seems reasonable that at least some of the Jacobians at each voxel will in practice be near or at the identity, so it appears that the bi-invariant mean will be only marginally more widely applicable than the log-Euclidean one. However, we will show below that there can be problems with the log-Euclidean *distance* between transformations far from the identity, even when they are close enough for the log-Euclidean *mean* to be well-defined. First, we consider the invariance properties in more detail.

In the case of SPD tensors, it is well-known that the log-Euclidean distance is only invariant to a matrix congruence with a geometric similarity transformation, as shown in section 4.2.7. For Jacobian matrices, a change of coordinates does not result in a congruence ($T \xrightarrow{x \to Ax} ATA^T$) but in a matrix similarity, $J \xrightarrow{x \to Ax} AJA^{-1}$. Just as for tensors, if we consider a geometric similarity $A = sR$ where $R^T = R^{-1}$, then the matrix similarity reduces to a congruence:

$$AJA^{-1} = sRJ(sR)^{-1} = sRJR^{-1}/s = RJR^T,$$

and so the distance is unchanged. For more general A, the distance will vary. However, Arsigny points out that the mean actually enjoys greater invariance properties than the distance upon which it is based [2]. This somewhat counter-intuitive result is actually quite trivial to prove, since the matrix logarithm of the log-Euclidean mean is linear in the matrix logarithms of the individual matrices, and the matrix similarity can be brought

outside of each individual log (equation B.8) giving (for any invertible A)

$$\begin{aligned}
\operatorname{logm}\left(\bar{J}\right) &= \frac{1}{N}\sum_{i=1}^{N}\operatorname{logm}\left(J_i\right) \\
\operatorname{logm}\left(\overline{AJA^{-1}}\right) &= \frac{1}{N}\sum_{i=1}^{N}\operatorname{logm}\left(AJ_iA^{-1}\right) \\
&= \frac{1}{N}A\left(\sum_{i=1}^{N}\operatorname{logm}\left(J_i\right)\right)A^{-1} \\
&= A\operatorname{logm}\left(\bar{J}\right)A^{-1}.
\end{aligned}$$

However, we note, with reference to Woods [1], that a change of template subject in a cross-sectional TBM study does not result in a similarity. Instead, one must consider transformations such as $J_i \rightarrow J_iJ_t^{-1}$, under which the log-Euclidean mean is not invariant. A similar point is made in [7] regarding diffeomorphisms, i.e. log-Euclidean analysis of diffeomorphic transformations as exponentiated constant velocity fields [8] is not invariant to the choice of subject.

Furthermore, with reference to Moakher [6], simple examples with rotations can show paradoxical results from both the log-Euclidean metric and its mean. For example, consider again rotations about the z-axis. Moakher's distance between a rotation of 10° and -10° is simply 20°,[5] and the average of these two rotations is the identity. If we rotate both rotations by 60°, so that they become 70° and 50°, then the bi-invariant distance of course remains at 20° and the mean moves to 60°. Interestingly, all these results also hold for the log-Euclidean case. This is only true for the special case of rotations around a common axis though, for which the ($2^{-1/2}$ scaled) log-Euclidean distance gives the same value as Moakher's distance. For more general rotations the two distances differ, as one would expect. For example, rotating the plus and minus 10° z-axis rotations by 60° around the x-axis results in the log-Euclidean distance between them increasing by just under one degree. More importantly, the results break down for large rotations; while the rotation by 60° around the z-axis showed the same results for bi-invariant and log-Euclidean distances and means, rotating by 180° results in the distance between the plus and minus 10° rotations (transformed to -170° and 170° respectively) becoming 340°. Informally, the log-Euclidean distance is measured the 'long way round', i.e. via that part of the great circle geodesic that avoids passing through the 'antipodal' point [6] or 'cut locus' [2] opposite the identity, not that part which is shortest. Similarly, the log-Euclidean mean of $R_z(d^\circ)R_z(10^\circ)$ and $R_z(d^\circ)R_z(-10^\circ)$ is given by $R_z(d^\circ)$ for $|d^\circ| < 170$, but by $R_z(180^\circ)R_z(d^\circ)$ for $170 < |d^\circ| \leq 180$. This problem of discontinuity in the mean under transformation might be avoidable by ensuring that rotations (or more generally, Jacobians) lie within a small enough region of the manifold around the identity, as suggested by Woods [1]. However, there is no avoiding the fact that the distance measure itself is

[5] The normalisation by $\sqrt{2}$ in Moakher's distance [6] achieves the intuitive result that two rotations around a common axis by $\theta_1$rad and $\theta_2$rad have a distance of $|\theta_2 - \theta_1|$. We report the results in degrees for convenience.

less appropriate for more widely-separated elements, even when they are technically 'close enough' to the identity. This point was raised in the context of rotations and more general transformations (with application to interpolation in computater animation) by Alexa [9], who proposed that a simple solution could be found. Unfortunately, this proposal has since been rejected by Bloom et al. [10]. The latter paper instead favoured a return to quaternion-based interpolation of rotations, however, this is of limited use for more general transformation matrices such as the Jacobian tensor.

5.1.2 An illustrative experiment

To further investigate the adequacy of the log-Euclidean metric on general affine transformations, a short Monte Carlo experiment can be performed. In section 4.2.9, the fractional and geodesic anisotropies were respectively related to the Euclidean and Riemannian distances between a general tensor and the closest isotropic tensor. Motivated by this, consider the following interesting question in the context of geometric transformations: can we find a purely rigid transformation that best approximates a more general affine transformation? First, note that the best translation in the approximating transformation would depend on the object, or region of space, to which the transformations are applied. In particular, given an optimality criterion of least squared error between landmark points that correspond under the affine transformation, the translation must be chosen so that the rigid transformation matches the centroids of the sets of landmarks [11]. This means that the essence of the problem is to find the closest rotation to a general linear transformation. The aforementioned lack of a bi-invariant Riemannian metric for matrices in $GL^+(3)$ is significant here; a pseudo-metric is of no use since the concept of 'closeness' requires a true distance metric.

Procrustes analysis — described in appendix C — provides a unique closed form solution for the rotation matrix that minimises the sum squared error between corresponding landmarks. Interestingly, like the translation, it turns out that the best rotation in this sense depends on the landmarks; there is no unique rotation that best approximates a linear transformation. We have not pursued a mathematical proof of this statement, but it is easily verified by simulating different sets of points and their linearly transformed corresponding sets, finding each time the rotation from the Procrustes method. Such repeated simulations allow us to compare the results of the optimal Procrustes fit (based on each set of corresponding points) to various other 'optimal' approximating rotations (based only on the affine transformation). For greater validity, we generate 50 random affine transformations (constrained to have no negative-real eigenvalues, and hence a real matrix logarithm), each of which is evaluated with 50 random sets of points. The number of pairs of points only needs to be two or more, to estimate a rotation in 3D, but we arbitrarily simulate ten pairs per set, to provide more accurate Procrustes fits. The point-sets allow the computation of the root-mean-square error (RMSE) for each of the rigid approximations, in addition to that of the RMSE-optimal Procrustes fit.

We argue that there are two reasonably obvious but somewhat ad hoc approaches for approximating a linear transformation with a rotation, without relying on an explicit

concept of distance. Based on the singular value decomposition (or alternatively, the polar decomposition presented in section 4.2.5) given that $L = USV^T = (USU^T)(UV^T)$, one candidate for an approximate rotation is $R = UV^T$. Another option derives from the fact that the Lie algebra for the Lie group of rotation matrices is the space of skew-symmetric matrices [6], or, in other words, rotation matrices have skew-symmetric matrix logarithms, $\operatorname{logm}(R) = K = -K^T$, and the rotation can be recovered from $\operatorname{expm}(K)$. Therefore, a rotation that approximates a linear transformation L could be generated from

$$K = \operatorname{logm}(L) \quad \text{(Not generally skew-symmetric)}$$
$$R = \operatorname{expm}\left(\frac{K - K^T}{2}\right).$$

There are also two obvious distance metrics: the basic Euclidean one afforded by the Frobenius norm of the difference, and the Riemannian log-Euclidean metric. Hence we may use numerical optimisation methods to solve

$$\underset{\mathrm{R}\in\mathrm{SO}(3)}{\arg\min}\ \mathrm{d}^2(\mathrm{R},\mathrm{L}) \tag{5.2}$$

for either $d_{\mathrm{Euc}}(R, L) = \|R - L\|_F$ or d_{LE}. In fact, since both metrics involve a sum of squares, efficient Levenberg-Marquardt algorithms [12] can be employed, such as that available in MATLAB's `lsqnonlin`.

Table 5.1 presents results from all five methods. Interestingly, the log-Euclidean 'closest rotation' exhibits worse performance than the naïve Euclidean one. This provides further fuel to the argument mentioned in section 4.3.6 (that log-Euclidean analysis has actually been found to be less powerful than using a Euclidean metric on diffusion tensor imaging data [13]) and additional motivation to investigate these issues, both in theory and in practice. Furthermore, the relatively good performance of the decomposition-based approximation could motivate a search for a suitable metric (or perhaps bi-invariant pseudo-metric) for which this is the closest rotation, providing a theoretical basis for this ad hoc approach.

	RMS Error		Rank RMSE	
Method	Mean	(Std)	Mean	(Std)
Procrustes	1.605	(0.668)	1	(0)
Decomposition	1.668	(0.667)	2.86	(0.95)
Lie algebra	1.918	(0.887)	3.82	(1.06)
Minimum d_{Euc}	1.694	(0.674)	3.31	(1.09)
Minimum d_{LE}	1.863	(0.805)	4.01	(1.00)

Table 5.1: Comparison of different methods to compute the 'closest' rotation to a given linear transformation. 50 different affine transformations were simulated (with no negative-real eigenvalues), and 50 sets of 10 corresponding points were computed for each one. The table reports the mean and standard deviation of the root-mean-square errors, and also the mean and standard deviations of the ranks of the different methods based on their RMS errors.

5.1.3 A theoretical connection and a practical compromise

We have described how Jacobian matrices or transformations far away from the identity can lead to problems with the log-Euclidean distance and consequently the log-Euclidean mean. On the other hand, the bi-invariant mean is suitable for almost all valid Jacobians, but cannot lead to second-order statistics without a Riemannian metric. It is therefore useful to try to combine the advantages of both approaches. If Woods' method is used to 'centre' each Jacobian matrix, replacing J_i with $J_i\bar{J}^{-1}$ using the bi-invariant mean $\bar{J}$, then the new bi-invariant mean of the centred matrices becomes the identity [1]. It is then tempting to consider the log-Euclidean statistics between these centred matrices, as a pragmatic compromise between full affine-invariance and the triangle inequality. In particular, we observe here that

$$d_{\mathrm{LE}}(J_i\bar{J}^{-1}, I) = \|\mathrm{logm}\left(J_i\bar{J}^{-1}\right) - \mathrm{logm}\,(I)\|_F = \|\mathrm{logm}\left(J_i\bar{J}^{-1}\right)\|_F, \tag{5.3}$$

which provides an interesting new justification for Woods' deviations, since their Frobenius norms are the log-Euclidean distances from the identity to the centred Jacobians. The distance between different centred Jacobians is then not $\|\mathrm{logm}\left(J_iJ_k^{-1}\right)\|_F$, but

$$d_{\mathrm{LE}}(J_i\bar{J}^{-1}, J_k\bar{J}^{-1}) = \|\mathrm{logm}\left(J_i\bar{J}^{-1}\right) - \mathrm{logm}\left(J_k\bar{J}^{-1}\right)\|_F. \tag{5.4}$$

Since the motivation for centring here is simply to bring the matrices closer to the identity, rather than to consider the Jacobians of the transformation from a hypothetical mean volume to the volumes under study (as in the original paper, which analysed cross-sectional data [1]), it would appear that the method is motivated even in the case of analysing longitudinal Jacobians.

5.1.4 Conclusion

In light of (a) the significant challenges, arising from the absence of an ideal metric and the open mathematical questions posed in the most recent work [2]; (b) the dearth of experimental results, notably, to the best of our knowledge, a complete absence of data analysed using Woods' method; and (c) the relatively large number of potentially competing or complementary methods available, and the number of possible extensions of them to serial imaging; it is clear that there is a major need for future research in this area, encompassing both theoretical investigation and experimental comparison of the different methods and their variations.

5.2 Differential Bias Correction

5.2.1 Introduction

Magnetic resonance images of spatially uniform objects typically exhibit intensity non-uniformity (INU). This arises from several different physical sources [14], and is typically assumed to be a smoothly varying multiplicative gain or bias field. It has been commonly observed that while INU appears to have little impact on expert interpretation of MRI (including neurological examinations), it can be of major importance in computational analysis, such as automatic brain-tissue segmentation or the quantification of longitudinal tissue loss [15, 16]. Segmentation methods based on intensity thresholds, or more complex fuzzy or probabilistic models of the intensity distribution are severely hindered by the spatial variation in signal intensity within homogeneous tissue, which blurs the intensity histogram and reduces the separability of the tissue classes. Techniques which attempt to measure volume loss through shifting boundaries using direct analysis of MR intensity [15] or edge detection [17] can be confounded by intensity differences due to inhomogeneity. A less frequently made point is that INU may also have detrimental impact on non-rigid registration: Studholme et al. [18] noted that voxel-wise maps of volume change from fluid registration can exhibit significant errors in the presence of local intensity variations, even when the registration is driven by an entropy-based similarity criterion such as normalised mutual information. Due to its importance, many approaches for the reduction or retrospective correction of INU have been proposed, as well as approaches which attempt to account for or reduce the impact of inhomogeneity within other procedures [18, 19]. Vovk et al. [20] provide a recent review, in which they distinguish between prospective and retrospective correction. Examples of prospective correction (or calibration) include the acquisition of uniform phantom images [21], the correction of highly non-uniform surface-coil data using more homogeneous body-coil images [22], and the use of modified pulse sequences (especially important at high-field [23]). Vovk et al. [20] focus on retrospective correction, which they further categorise into methods based on spatial filtering [24], fitting surfaces to data-points [25], segmentation models [26, 27], and on the intensity histogram [28]. The review by Vovk et al. [20] also includes a discussion of metrics for performance evaluation, and a comprehensive survey of publications and available software, but no direct comparison of methods is presented. The most complete qualitative and quantitative evaluation of retrospective bias correction algorithms appears to be Arnold et al. [29], which compared six algorithms on real and simulated data, and found the locally adaptive methods N3 [28] and BFC [30] to be the most successful.

Application to serial imaging

For longitudinal analysis of MRI, it is the differential intensity non-uniformity between the two (or more) images of a single subject which is of greatest importance. The boundary shift integral (BSI) [15], SIENA [17], and serial rigid or non-rigid registration algorithms driven by intensity differences or correlations [31, 32, 33], should be insensitive to the

component of the two images' bias fields which is common to them both.[6] For this reason, the focus here will be on methods for differential bias correction (DBC), proposed by Lewis and Fox [16]. Little research has been reported specifically on this topic, though very relevant methodological work has been published for related techniques, such as multi-image and template-based non-uniformity correction schemes (reviewed below). The central problem in DBC is separating intensity differences due to the bias field (assumed to be relatively smoothly varying) from those due to noise or brain volume change, which will generally be of much higher spatial frequency. Greater amounts of brain atrophy make the differential bias harder to estimate due to larger areas with unaligned tissue boundaries, and lower spatial frequency of the resulting atrophic component. In this work, a new approach of integrating DBC with intra-subject non-rigid registration is proposed; the key motivation being that even partial removal of differential bias should improve non-rigid registration, and similarly, even incomplete image alignment will greatly enhance DBC, as it will dramatically reduce the intensity differences from originally unregistered tissue boundaries.

5.2.2 Background

Physics and nature of inhomogeneity

There are a number of complex physical processes underlying the observed intensity non-uniformity [14, 23, 28, 34, 35]. In this brief review, the focus will be on the magnitude and character of the non-uniformity, particularly those aspects which may be object-dependent, or less consistent with the typical assumptions. Effects are more severe (and harder to characterise) at higher field strengths. Since the aim of this work is to develop improved correction methods, which in turn require means of evaluation, particular emphasis is placed on the use of simulation techniques in the literature.

Sled et al. [28] listed the following basic physical sources of intensity non-uniformity:

- frequency response of the receiver
- spatial sensitivity profile of unloaded receive coil
- induced currents, standing wave effects (also known as dielectric resonance, but now more properly understood as field focusing)
- excitation field inhomogeneity

With surface coils, the effect of spatially variable receive coil sensitivity is dramatic and dominates the other sources (as utilised in some multi-image correction approaches discussed later). Head coils also show signal reductions (though much less dramatic) toward the boundaries of the coil [36]. Some intensity variation can be caused by geometric distortion; Wang and Doddrell investigate the characterisation and correction of this using phantom-based methods [37].

[6] Strictly, the BSI is invariant to the arithmetic average bias field, but not the geometric average, which may be considered more natural, given the positive multiplicative nature of INU.

Induced currents, and RF focusing effects depend on the electrical and geometric properties of the imaged object, as well as the pulse sequence and coil polarisation [34]. These effects are less consistent with the common assumption of a smoothly varying multiplicative field, and may also produce tissue-dependent contrast changes. Simmons et al. [14] considered 'standing wave' effects and RF penetration or skin-depth phenomena; they found these to be negligible in the human head at 1.5T, but other authors have since discovered such sources of inhomogeneity to be more significant at higher field strengths. At 3T or above, intensity non-uniformity is particularly problematic [23, 35, 38]. Cohen et al. [38] reported that inversion-prepared low flip angle (FLASH-type) sequences seem to be especially sensitive. Both field inhomogeneity and susceptibility artefacts increase with field strength, and above 1.5 T it is not possible to neglect RF eddy currents induced in the body, or 'standing wave' effects along the sample [35]. RF focusing effects exacerbate B1 inhomogeneity at high field strengths; Thomas et al. [23] develop a modified pulse sequence to reduce problems in T1-weighted MDEFT imaging at 4.7 T. For T2-weighted imaging, fast spin echo sequences can be relatively insensitive to inhomogeneity, due to coherency in partial echoes [39].

Magnitude and scale of intensity non-uniformity

Sled et al. [28] used manually fitted bias fields (the model fitted was not specified) to investigate the typical properties of non-uniformity in twelve individuals, each scanned on a different MR machine. Scanner manufacturers included Philips, Siemens, and GE. The paper does not indicate the field strength(s) of these scanners, but 1.5T seems likely for research institutions in the mid-nineties. Histograms were presented for non-uniformity field strength, showing approximately unimodal distributions with FWHM ranging from about 10% to 20% for 3D gradient echo T1 volumes, up to around 40% for T2 images coming from a multi-slice dual echo (T2/PD) spin echo sequence. Alecci et al. [35] investigated the relative magnitude of the B1 field using numerical simulation and measurement in phantoms and volunteers, imaging with a birdcage coil at 3T. They found variation of 15% across the brain in the transverse plane, and large variation along the z-axis of the coil (less than 10% over 7cm, but rising to around 35% over 15cm). Little research seems to have been reported on the typical spatial frequency characteristics of the bias field.

Inhomogeneity simulation

Controlled generation of artificial intensity non-uniformity is very useful for evaluation (see below) but has typically been quite simplistic to date, with little attempt to model either the underlying physics or the complexities of the observed phenomenon. As Sled et al. [28] state, characterising the physics is difficult. Balac and Chupin [40] have developed numerical methods for determining object-dependent RF inhomogeneity artefacts, which they propose to implement in an advanced MR simulator [41], but this work has not yet been used for comparison or refinement of bias correction algorithms. Following investigation of the nature manually extracted real bias fields, Sled et al. [28] simulated two non-uniformity fields using combinations of quadratic and Gaussian terms, varying

in magnitude by 20% within the brain volume (similar to their observed FWHM of 10–20% for T1 volumes). It appears from figures in the paper that the fields are intended to exhibit typical patterns of diagonal variation (see [34]) and field-focussing 'hot-spots' [23], though this is not explicitly stated. The extremely popular MNI BrainWeb [42] has its simulated INU fields available for download. These were derived from actual scans, so are realistically complex, but little is said about their extraction or parametrisation. They have been subsequently normalised to exhibit the same 20% within-brain variation that Sled simulated [28]. Note that by multiplying the fields and then offsetting them to restore the unity gain level, it is possible to generate arbitrary levels of inhomogeneity, though such simplistic adjustment will obviously not account for the changes in the nature of the inhomogeneity at higher field strength. Note too, that if such fields are applied to different subjects (especially scans with different fields of view or orientation), they may no longer have realistic relationships to the subject geometry or other properties. Arnold et al. [29] simulated two different types of fields. The first used 'orthogonal' polynomials (presumably orthogonal in terms of being independent in the three dimensions, with no cross terms; rather than anything to do with Legendre polynomials or similar), and the second was a sum of orthogonal sinusoids, with periods from 0.8 to 1.2 times the field of view. Both fields were generated with three magnitudes of 4, 8 and 16% total variation (apparently over the entire field of view (FOV), not the brain region). These magnitudes are obviously smaller than those proposed by Sled et al. [28], or offered by default in BrainWeb; no motivation from either physics or empirical data is given for this choice. Arnold et al. [29] do not clarify whether these simulated fields are intended to partially account for the geometry or other properties of subject. Both a simulated field and a phantom-measured prototype were applied to simulated brain phantoms in Brinkmann et al. [24]. Their simulated field was a simple linear ramp which they created in versions with 22% and 96% variation over the FOV. The phantom-measured field came from phased array surface coil imaging of a uniform water phantom; it originally exhibited variation of 329%, after which they created three reduced versions (186%, 38%, 10%). Clearly neither field related to the (phantom) brain properties.

5.2.3 Correction methods

Simmons et al. [14] concluded their investigation of the physics of intensity non-uniformity at 1.5 T by stating that the use of uniform oil phantoms is 'generally more appropriate than correction based on low pass filtering'. However, other authors have pointed out that phantom-based methods are unable to correct for object-dependent inhomogeneities, which may be significant [43]. Particularly at higher field strengths, RF-focusing and other effects motivate the use of retrospective correction algorithms. In addition, automatic methods for bias correction have increased in theoretical sophistication and computational capabilities since the 1994 publication of Simmons et al. [14]. Techniques for non-uniformity correction are too numerous to review comprehensively here; instead, a few notable algorithms are discussed, with emphasis on the phenomenological models they employ, and a particular focus on methods developed for multi-image situations.

Sled et al.'s [28] nonparametric nonuniform intensity normalisation algorithm (N3) appears to be the most popular bias correction software [20], and one of the most successful approaches [29]. It is based on the premise that the presence of the bias field blurs the intensity histogram, as pure tissue peaks are spread out by the range of their non-uniformity. N3 sharpens the histogram by iteratively deconvolving narrow Gaussian distributions, smoothing the estimates between each iteration by fitting a cubic B-spline (200mm spline support, corresponding to 50mm knot-spacing). The algorithm does not require a tissue model, nor the identification of pure tissue regions. N3 has also been studied with particular reference to the imaging of AD [44]. The bias field corrector (BFC) [30] attempts to match local histograms to the overall intensity distribution, again fitting a cubic B-spline with 64mm knot-spacing. Ashburner and Friston [27] introduced a unified approach that combines spatial normalisation, tissue segmentation, and bias correction into a single generative model, implemented in the SPM5 software. They parametrise the bias field with a Fourier (Discrete Cosine Transform) basis, regularised following a model of smoothed Gaussian noise. The model's highest frequency basis function has a period of 60mm by default. No comparisons of SPM5's bias-correction performance relative to SPM2 and/or other bias correction techniques seem to have been published thus far.

Multi-image techniques

Bromiley and Thacker [45] developed the method of Vokurka et al. [46] into a specialised algorithm for situations in which it is possible to acquire multiple images of the same object with approximately the same INU but differing tissue contrast. The main case in which this approximation may reasonably be assumed is surface coil imaging, since the overriding inhomogeneity of the receive coil sensitivity is largely independent of the changes in pulse sequence needed to alter image contrast. Lai & Fang focus exclusively on surface coil images; their correction technique uses a lower resolution body-coil image which they treat as being free from inhomogeneity for the purpose of correcting the far greater non-uniformity of the surface coil image. They align the two images with a rigid+scaling transformation model, then select points for which the surrounding region can be well approximated by a plane, and then fit a membrane spline model to the ratio of surface- to body-coil images at these points using a preconditioned conjugate gradient optimiser. The complete algorithm appears to be very fast, though the authors only report results in 2D on relatively small images (with a 64×64 body-coil acquisition and a 128×128 surface-coil image, run time is only a few seconds). The paper appears not to mention the typical spacing or number of the points used to fit the spline, though the figures show extracted bias fields that appear relatively complex, compared to low-order polynomials or Fourier bases. An interesting and unusual multi-image approach has been developed by Learned-Miller and Jain [47], they consider a set of images of different subjects (around 20), and reduce the bias from each down to the set's shared common component. In stark contrast to methods that assume homogeneous tissue regions can be found, they treat each voxel as being independent; instead considering the entropy of each voxel's intensity over the set of images. They note that minimising the sum of all these voxel-wise entropies

will tend to remove intensity variations not shared by all images. Their method requires the images to be approximately, but not perfectly, aligned — misregistration will increase the entropy over the image set, but they assume that additional intensity non-uniformity would further increase entropy. The inhomogeneity field is modelled with a DCT basis, with shortest period equal to half the FOV. Experiments on a set of infant brain scans show impressive results, with large inhomogeneities apparently removed, while genuine developmental differences in white matter intensity remain.

Differential bias correction

The particular multi-image problem focused on here is the correction of differential bias in longitudinal MRI data. The original DBC algorithm developed by Lewis and Fox [16] uses median filtering of the difference image from log-transformed originals (equivalent to filtering the log of the image ratio). The exponential of the processed result is then applied equally to baseline and repeat images (multiplying by the square root and its reciprocal respectively). Smoothing is performed with an $11 \times 11 \times 11$ voxel box kernel median filter, carried out over a particular region, with zeros outside the region being included for voxels near the border. The region is created by dilating (with a 6-connected 3D cross structuring element) the binary union of the baseline and repeat brain regions. As a preprocessing step, the original images are first normalised to have the same mean intensity over their respective brain regions (justifying the inclusion of zeros when filtering the log ratio image). A related approach has been implemented by Gaser[7] for application to longitudinal voxel-based morphometry. Gaser applies a 30mm FWHM Gaussian to smooth the difference image (in original, not log-space) over the intracranial region, and then uses this to correct the image pair. Methodologically, the most closely related paper to the work discussed here is actually a single-image bias correction method [48], which uses non-rigid registration of the image to a manually bias-corrected template image, allowing differential correction to remove the bias in the original source image. Studholme et al. [48] implemented this approach by filtering the original and registered template images separately, before taking the ratio. Smoothing is performed over a template-space brain mask using a Gaussian kernel, including re-normalisation of the kernel to account for missing data outside of the mask. Comparison to manual correction showed that a kernel of only 20mm FWHM gave the best results, in surprising contrast to the common assumption of a much smoother bias field, and the empirical results of Brinkmann et al. [24] mentioned below. The non-rigid registration algorithm uses a B-Spline Free-Form Deformation model, with multi-level and multi-resolution optimisation at knot-spacings halving from 14.4 to 1.8mm, and Gaussian blurring with FWHM of 4, 2.4, 1.2 and 0.6mm. A bending energy term is used to regularise the transformation. Studholme et al. [48] optimised their algorithm specifically for elderly subjects or those with neurodegenerative disease.

[7] `http://dbm.neuro.uni-jena.de/vbm/vbm2-for-spm2/longitudinal-data/`

Relevant research on filtering

While not addressing the DBC problem itself, there is useful literature related to the smoothing or filtering used in the above-mentioned differential techniques. Implementation choices such as the filtering method are likely to be of importance in DBC performance [16]. Several simple, fast, and popular techniques for single-image bias correction involve the use of a low-pass filtered version of the image to correct the original version. One of the earliest such methods is known as homomorphic unsharp masking (HUM) [49], which, in essence, simply divides the original by its smoothed version. HUM has been referred to as an approximation to true homomorphic filtering [24], in which the convolution is performed in log-space [50]. Brinkmann et al. [24] investigated the use of either mean or median filtering for HUM, at a range of different spatial scales. They worked in 2D, using only square box kernels, with voxel-dimensions from 5×5 to 383×383. They found that the mean almost invariably outperformed the median filter (something worthy of further investigation with DBC), and that much larger kernels (e.g. 64×64) outperformed those reported in past literature. Working in 3D, and on 3T images, Cohen et al. [38] used efficient FFT-based Gaussian convolution, after filling background voxels with the foreground mean. Their kernel size was chosen in proportion to the image field of view, having a 'half-width' of 3/8 times the FOV in voxels.

5.2.4 Evaluation methods

Vovk et al. [20] review evaluation in depth, discussing: qualitative analysis; intensity based statistics, such as coefficient of variation within tissue classes, or coefficient of joint variation between two classes; indirect measures of bias correction performance via subsequent segmentation results; and quantitative comparison of corrected images or extracted bias fields, using either simulated or carefully measured 'gold-standard' data. The latter seem the most direct quantitative measure, though gold-standard data is difficult to obtain validly. Oil or water phantoms (as used e.g. by [14, 24]) may not relate well to real brain images. Painstakingly manually corrected data such as that used by Studholme et al. [48] has several advantages, but may introduce potential bias in terms of matching of the evaluated models to that used for the gold-standard. Reliability of the manual estimation is also likely to be spatially varying, due to variable ease of selecting regions of pure tissue. Simulated bias fields are attractive, in that the ground truth is precisely known, and the parametrisation of the gold standard can be easily varied independently from the correction models being evaluated. If simulated fields are applied to real scans, the overall inhomogeneity is then unknown, which has motivated the popular application of simulated bias to simulated images [24, 28, 29]. Note however that for DBC, a single real image with an added simulated field has a known differential bias, regardless of the inhomogeneity of the original scan.

Metrics for quantitative evaluation

Sled et al. [28] suggest that the best evaluation approach is to compare the estimated and true field, if the latter is known. They note that constant scale-factor differences between bias fields are of no interest; motivated by this the metric they use is the coefficient of variation of the ratio of estimated to the actual field. Studholme et al. [48] measure the RMS error between their manually corrected gold standard intensities and the output images from the automatic methods, computed over a manually defined intracranial region. The images are globally rescaled to match their within-brain mean intensities, due to the fact that constant overall scaling of the intensities is irrelevant. They also report the coefficient of variation of intensities within manually identified regions of supposedly pure white matter, with the view that all variation within this region comes from the bias field. This assumption is slightly inconsistent with their later work [18], which highlights the potential for genuine variability in tissue intensity (for example due to development, aging or disease), but the metric nevertheless provides useful additional information. Shattuck et al. [30] also consider the RMS error, between the unbiased BrainWeb phantom and the corrected images; they suggest an affine or Procrustes transformation to match the original and corrected intensities, on the basis that constant intensity offsets would not affect their subsequent automatic segmentation process. Note that measuring errors in the corrected intensities instead of the recovered non-uniformity field, effectively weights the bias field errors by the tissue intensity — the same magnitude of error in the field estimate will be more significant within white matter than in grey matter (on T1 scans). This might give an overly favourable impression of methods which fit models to segmented white matter or selected points within it, compared to those which do not distinguish between tissue types, though the effect is probably minor. On the other hand, correction of non-uniformity in regions with very low signal intensity may genuinely be less important, as these are often not the focus of further analysis. Note also that estimation of the bias field is inherently less reliable in low intensity regions; there is less information in the quantised MR signal, and the Rician distributed noise [51] is more heavily skewed.

An interesting complement to the direct measures thus far described is the stability analysis of Arnold et al. [29]. They observe that an ideal bias correction technique would not only remove most of the bias in a single application, but that repeated applications of the technique should not continue to have major effects on the corrected volume. They discover that HUM fails to meet this desideratum.

5.2.5 Preliminary investigation

Introduction

In this section, a new approach for combined longitudinal bias correction and registration is described, and a pilot study to explore the technique's potential is reported. The method interleaves the DBC method of Lewis & Fox [16] between the stages of a multi-level B-Spline FFD non-rigid registration algorithm [52, 53]. The increasingly precise alignment should allow progressively better bias correction performance, which in turn,

should improve the accuracy and validity of the subsequent registration steps.

Method

At this stage, the DBC procedure has been reimplemented closely based on that from Lewis and Fox [16]. The only modification being the use of a spherical kernel with variable radius in place of the cubic kernel with 11 voxel side-length. Since the amount of unregistered voxels that need filtering out should reduce with each registration stage, the kernel size is decreased at each iteration. It would be possible to initiate the iteration with either non-rigid registration or bias correction first; here, the images used for testing are rigidly registered to begin with, so it was decided to start with a bias correction step. A B-Spline FFD non-rigid registration algorithm is used [52], with Normalised Mutual Information [54] as the similarity function. Based on the work of Boyes et al. [55] on longitudinal B-Spline registration, the sequence of control-point spacings used in the multi-level optimisation is 10, 5, and 2.5mm. Following the finest FFD level, a final DBC step is performed. The five DBC steps use kernel radii of 7, 6, 5, 4, and 3 voxels, where the first of these corresponds to a volume slightly larger than the 11-voxel cubic kernel used by Lewis and Fox [16].[8] In the first four DBC steps, the reciprocal of the estimated differential bias field is applied entirely to a copy of the target; the original source image is then re-registered to this copy, with the aim of avoiding the accumulation of bias field errors in the imperfectly registered source. After the final step, the estimated bias field could be split between the target and transformed source as in the original method of Lewis and Fox [16], with the slight complication that the inverse of the non-rigid registration would be required to map the bias field to the untransformed source space. Methods for the direct inversion of B-Spline FFDs seem not to have been published; dense voxel-wise fields can be inverted [56, 57] and then re-approximated with B-Splines [58].

The method has been tested on three pairs of images with simulated Alzheimer-type atrophy. The baseline images are real images of patients with probable AD, and the repeat scans are generated from these to mirror the subjects' real measurements of whole-brain and hippocampal volume loss over a period of one year, using the thermoelastic phenomenological model of Camara et al. [59]. Amounts of simulated brain atrophy for the three subjects were: 3.15, 1.47, 2.05 per cent of original whole brain volume. Because the repeat scans are simply deformed versions of the original, there is no true differential bias (though there will initially be some due to misregistration of the original bias present). Simulated bias was applied using a version of the BrainWeb 20% INU field ('rf20_A'). The field image was geometrically scaled along each axis to cover the FOV of each patient image and resampled using linear interpolation (the resultant slight blurring is of no concern since the field contains only low spatial frequencies). Following the simulation of atrophy and the addition of the multiplicative bias field, Rician noise is simulated in the baseline and

[8] A 7 mm sphere has volume $4\pi r^3/3 = 1436.8\,\mathrm{mm}^3$, and a discretised version contains 1419 voxels; an 11-voxel cube contains 1331 voxels.

repeat images by adding independent Gaussian variates to each voxel in quadrature [28],

$$s \to \sqrt{(s+n_1)^2 + n_2^2}, \qquad where \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} \sim \mathrm{N}(0, \sigma^2 I).$$

The Gaussian standard deviation was set proportional to a datum of the 95th percentile of image intensities, as this should be more robust than the potentially unreliable maximum intensity (e.g. in the presence of a few unnaturally bright voxels from pulse artefact). Standard deviations of 1 and 1.5% of datum were used for the baseline and simulated repeat images respectively, with greater noise in the repeat in an attempt to compensate for the interpolation-based smoothing of the original noise present in the baseline when this is transformed into the repeat image. Because of the simulated nature of the atrophy, bias field, and noise, ground truth results are available for comparison with the estimated deformation and bias fields. The results can be quantified over either the original baseline subjects' semi-automatic brain tissue segmentations, or over propagated BrainWeb intracranial volume labels, used in the atrophy simulation meshing process.

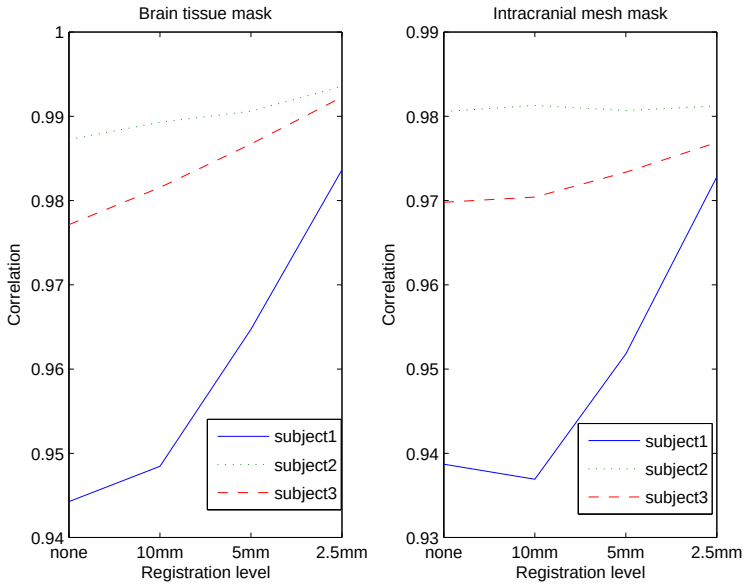


Figure 5.1: Pearson correlation of estimated with true bias field after each registration step. The metric has been evaluated over a tight brain-mask and a more inclusive intracranial region.

Results

Figure 5.1 shows bias correction results for the three subjects, before registration, and after each of the three registration levels, with both masks, in terms of the correlation of estimated and true bias fields. Figure 5.2 shows two further metrics: coefficient of variation of the ratio (as used by Sled et al. [28]) over the brain mask, and RMS Error between estimated and true fields over the intracranial mask (note that no scaling has been deemed necessary here, since DBC does not add arbitrary multiplicative constants).

Results have also been produced for the remaining combinations of mask and metric; these are very similar, and are not shown here.

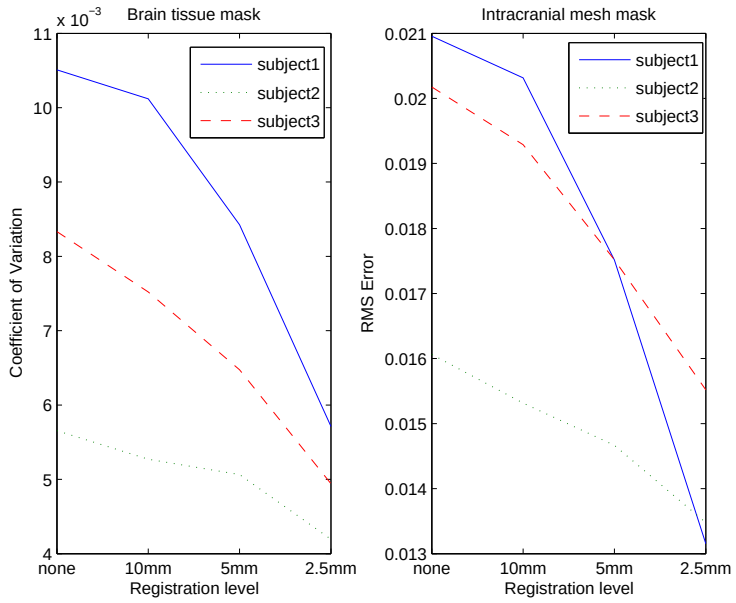


Figure 5.2: Coefficient of variation (std/mean) of the ratio of estimated to true bias field, evaluated over the brain mask, and RMS error between estimated and true fields over the intracranial mask.

Figure 5.3 presents registration results in terms of average Euclidean distance between estimated and true displacements, evaluated over the intracranial mesh, for the three registration levels, both with (left) and without (right) interleaved differential bias correction. Figure 5.4 gives similar results quantified in terms of the RMS error in log-transformed Jacobian determinants, over the brain tissue mask. Displacement fields evaluated over the brain tissue, and Jacobians compared over the intracranial mesh exhibit similar patterns.

Discussion

At this stage, with only a small number of subjects analysed, the results should not be interpreted too strongly; they are, nonetheless, very encouraging. Regardless of the exact metric used to quantify the bias correction performance, or the mask region over which it is analysed, results for all three subjects show an almost universally monotonic increase in bias correction performance with more precise registration. In all cases, the final bias correction is significantly better than the original use of DBC on the images with only rigid registration.

At each level of registration, the addition of interleaved DBC seems to improve the accuracy of the registration results (compared the FEM gold-standard). There appears to be some evidence that without DBC (right-hand plots of figures 5.3 & 5.4) the registration accuracy degrades with finer control-point spacings, as the algorithm begins to register the bias (an example of over-fitting, since the similarity measure is increasing all the time). With DBC however, the accuracy (as well as the NMI objective function) increases with

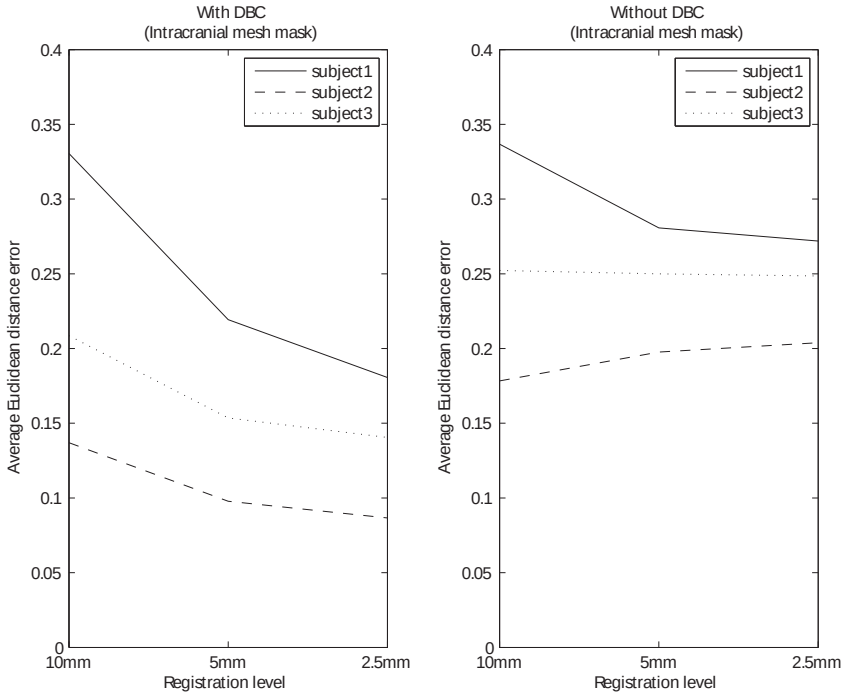


Figure 5.3: Error in estimated deformations (Euclidean distance between estimated and true displacement vectors), evaluated over the intracranial mask. Results are shown for each registration level, with (left) and without (right) differential bias correction.

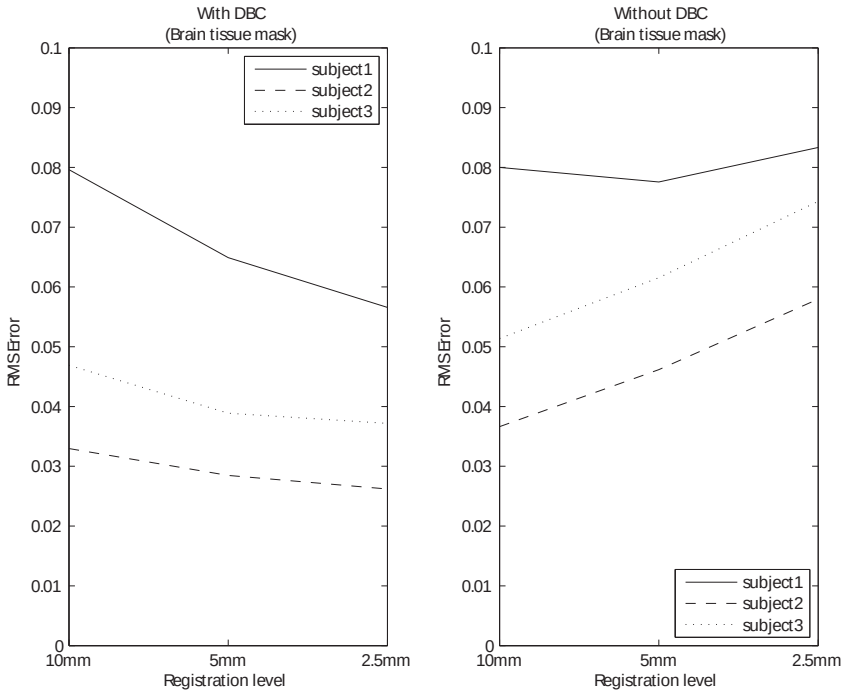


Figure 5.4: Error in volume change maps (RMS error in $\log|Jacobian|$) after each registration level, evaluated over the brain mask. Left: with DBC; right: without.

every finer level of registration, whether measured via direct displacement field agreement or Jacobian determinants. This suggests that registration in the presence of bias might be genuinely improved through the use of DBC, in terms of more accurate correspondence, rather than just better image similarity. See Crum et al. [60] for detailed discussion regarding registration performance. Final NMI image similarities for the three subjects are given in table 5.2, showing that DBC also significantly improved image similarity — even though NMI should be relatively insensitive to bias, compared to simpler measures such as summed squared error or intensity correlation.

Method	subject 1	subject 2	subject 3
with DBC	1.72805	1.77095	1.74017
without	1.62507	1.64478	1.63176

Table 5.2: Final values of the Normalised Mutual Information similarity criterion, at the finest level of the FFD registration, with and without the interleaved use of differential bias correction.

Interestingly, in their work on template-based single-image bias correction, in which a single multi-level registration step is followed by a single bias correction step, Studholme et al. [48] commented that further registration and bias correction appeared to be generally unrequired. However, the crucially important distinction here is that they did not have access to gold standard registration results; only their bias field gold standard. So their comment may simply indicate that iterated registration and bias correction did not improve the similarity criterion in their case — the registration accuracy may actually still improve. It is possible to get gold standard registration results for inter-subject correspondence, for example, via the simulation of deformation fields from population statistics [61], so this could be investigated more carefully.

Another interesting comment made by Studholme et al. [48] is that B-Spline models with energy-based regularisation terms (cf. the volume preservation terms used in [62, 63]) may wrongly attribute volume changes to volume-preserving displacements. Here, we have found the multi-level B-Spline model [53] sufficiently robust to require no such regularisation term, though this question may be worthy of further investigation.

5.2.6 Plan of future work

Further research on this topic will involve both modification of the techniques, and much more thorough optimisation and evaluation of the results. The combination of non-rigid registration and differential bias correction opens up a bewildering array of options regarding the exact methodology, since both registration parameters (such as the choice of control-point spacings) and bias correction parameters (such as kernel size) are very likely to interact. Thanks to the possibility of simulating bias fields and quantitatively measuring performance, it may be feasible to do much of the parameter tuning automatically, using a powerful computing cluster to perform numerous runs of the algorithm. However, simulated data has its limitations, so manual qualitative assessment of the effect on real images will also be necessary.

The specific objectives of further work should be to firstly determine good, if not perfect, settings for the existing methodology, then to demonstrate through careful evaluation that the technique offers sufficient benefits to either (ideally both) bias correction or non-rigid registration of typical brain images to be worth the extra computing resources required. Following that stage, more novel methodological development could be considered, for which some initial ideas are presented next.

5.2.7 Methodological development

Modification of the existing technique

The most obvious area for potential improvement is in the choice of smoothing method. The original paper on differential bias correction [16] speculated that frequency domain filtering could be more successful, as it can explicitly separate different components of the spatial frequency spectrum, with the possibility to distinguish bias from atrophy, registration error, and noise. Filtering can be done either in log-space (homomorphic filtering [50]) as used by Lewis and Fox [16], or in the original intensity space (as for homomorphic unsharp masking [24, 49]). There is also a decision to be made whether the ratio of the original images is filtered, as in Lewis and Fox [16], or whether the ratio is taken after filtering the originals, as in Studholme et al. [48]. The size of the filtering kernel can be easily adjusted, though finding theoretical support for the choice is harder. The original DBC paper states that the $11 \times 11 \times 11$ voxel cube was chosen on the qualitative basis of apparent removal of anatomical structure. The work of Brinkmann et al. [24] on single-image correction suggested kernels should be larger than typically used. However, the problem of filtering the differential bias is clearly different to that of smoothing the original image; only the moving boundaries of atrophic regions need to be smoothed out, rather than the main structure of the brain. Brinkmann et al. [24] found that large (64×64 voxel) mean filters performed best for the HUM examples they investigated. They worked only in 2D though, and it seems conceivable that the use of 3D filtering might alter the optimal kernel size, since many more voxels are included in cubes of equivalent side-length. Disappointingly, they also neglected to compare HUM with homomorphic filtering, despite commenting that HUM is an approximation to it (the 'unapproximated' convolution in log-space is by no means computationally demanding). So there is clearly scope for further research here, both in the single image case and for differential bias correction. The other main option if convolution filtering is used is whether to use the median, mean, or something else. Lewis and Fox [16] used the median filter due to its desirable characteristics of removing Gaussian noise (the Rician distribution of MRI magnitude data is approximately Gaussian in foreground voxels) and erasing structures smaller than the kernel radius. Studholme et al. [48] used Gaussian (weighted mean) filtering in their template-based bias correction, but without theoretical justification. Considering Brinkmann et al.'s finding that median filtering is sub-optimal [24] (perhaps due to being constrained to choosing its output from among the input values), but in light of the obvious problems in differential bias correction of outlying atrophic voxels biasing the mean, an appealing alternative might be the

investigation of robust statistics such as the trimmed mean [64, 65]. A trimmed-mean filter could reject outlying intensity differences from atrophy, while maintaining some of the mean filters apparently superior bias field estimation properties. A number of other robust statistics may be worth investigation [66, 67].

Another important aspect with convolution filter smoothing is masking. Lewis and Fox [16] mask the log ratio-image (as described earlier) before filtering it; for boundary voxels, neighbouring zero-voxels outside the mask are included in the computation of the median. Including zeros has the effect of pulling the estimated bias field toward unity (after the anti-log transform), which is unlikely to be harmful, but may be worse than treating data outside the mask as being censored, and computing the median only of the valid data. Studholme et al. [48] use masking with Gaussian filtering, before dividing by the filtered mask image, which is equivalent to re-normalising the truncated Gaussian kernel near the boundaries to have unit sum. This also allows the estimate to be extrapolated beyond the original mask, and possibly beyond the region for which it is reasonable. Borrowing the mathematical terms domain and range, the domain mask could be defined to include the set of input voxels which can be considered to provide useful information for the estimate, while the range mask (which could be smaller or larger) would define the set of voxels for which an output is desired and/or for which the extrapolation may be trusted. At this stage, an intuitively appealing idea is that the domain mask should be the intersection of the two brain masks, while the range mask would cover at least their union, probably dilated. Outside the range mask the estimate of a unity gain field seems most reasonable. Note also that masking choices exist with the registration algorithm too. In the work reported above masking was not used for registration, but doing so may improve the results by ensuring that irrelevant extracranial changes cannot influence brain structures. The B-Spline FFD algorithm of Schnabel et al. [53] allows for voxels outside a target-space mask to be ignored in the similarity criterion, and/or for masked-out control-points to be made passive. The use of a source-space mask would require minor changes to the code. On the other hand, the procedure used to generate the above results included transformation, B-Spline interpolation, and subsequent re-binarisation of the source mask after each registration level, so that both target and source masks could be used for the DBC step, which may be an unnecessary computational burden (especially with shrewd use of separate domain and range masking).

Model-fitting

As an alternative to convolution (or FFT) based filtering methods, smoothing can also be effected by fitting parametric models to the data — often fitted directly to the intensity non-uniformity (see e.g. [25]; Vovk et al. [20] mention many others), here they could be used to model the differential bias field. B-Spline models are used in the popular and high-performing N3 [28] and BFC algorithms [30]. Note that control-point spacings for bias field models tend to be much higher (of the order of 50 or 60mm [28, 30]) than those used for non-rigid registration (typically below 20 mm and sometimes as low as 1.8 mm, [48]), though the manually corrected gold standard data of Studholme et al. [48] utilised

a B-Spline with 10mm control-point spacing. It seems possible that Studholme et al. [48] might have over-corrected their gold standard, removing genuine anatomical tissue variation, and that their subsequent finding of small kernels to be optimal could simply reflect this, but much more work (ideally grounded in the theory of MR physics) would be needed to support this speculation.

Using a model-fitting technique instead of filtering allows an alternative (possibly more principled) approach to outlier detection and removal. Instead of using median filtering or robust statistics to remove extreme values, the goodness of fit can identify, and either remove or down-weight, outliers to the model. Techniques such as iteratively-reweighted least squares [68], the robust mixture-model based GLM implemented as an extra in SPM5, or the RANSAC algorithm popular in computer vision [69], might all be worth consideration. A related idea was used in the dual image approach of Lai and Fang [22], for the correction of a surface-coil image using a body-coil reference. They fit a membrane spline to points in the ratio-image which are 'locally predictable' (based on the size of the error from fitting a plane within a local window around each point). This concept may also be useful for longitudinal differential correction, since atrophic and artefactual changes other than INU are likely to be less locally predictable than the more smoothly varying bias field.

DBC's relation to registration

There is also the potential to incorporate iterative differential bias correction within a fluid-registration algorithm [70]. Large-deformation techniques typically include multiple re-gridding steps [32] which involve the generation of a new source image (and the accumulation of displacement fields); at each of these re-griddings it would be possible to interleave either the simple filter-based DBC procedure used thus far, or a more complex model-based algorithm. Interestingly, however, Studholme et al. [48] advise against the use of such highly localised registration techniques in the presence of regional tissue variations. Instead, they went on to develop a new regionally-computed variant of mutual information [18], this RMI similarity criterion also aims to cope with true local tissue intensity variability arising from neurodegenerative disease or developmental processes – effects which may be of too fine a scale for common retrospective intensity correction methods to model. It may be important to consider such an approach, and possibly other further removed alternatives, such as feature-based non-rigid registration techniques like those of Xue et al. [71] and Cachier et al. [72]. Similar to the work of Studholme et al. [18],[9] Loeckx et al. [73] develop a measure of conditional mutual information (CMI), in which the intensity-based MI is essentially computed locally from conditional probabilities within spatial 'Parzen window' kernels.

The most ambitious goal for future work would be the development of a fully integrated — rather than just interleaved — combined longitudinal non-rigid registration and bias correction. Like the unified segmentation model of Ashburner and Friston [27], a single

[9] In fact, it appears equivalent, but correspondence with Loeckx reveals that this appearance is due to an error in [18].

generative model could be optimised (albeit perhaps with an Iterated Conditional Modes algorithm as they used), but in this case tailored to the analysis of serial MRI. There is limited work in the literature on combining bias correction and registration, the most notable work appears to be that of Knops et al. [74], who show that the incorporation of a bias correction method reduces the number of misregistrations without loss of accuracy in already-successful ones, and some currently unpublished developments by Andersson et al., which are available in the software tool `fnirt` within FSL.[10] The review by Vovk et al. [20] contains a suggestion that combined information theoretic registration and bias correction should be explored. Lewis and Fox [16] also mentioned the possible use of joint intensity histogram sharpening as a bias correction optimisation criterion. Since joint entropy based methods have proven successful for non-rigid registration, the use of a single mutual information (or similar) objective function for both tasks is very appealing, and, apparently, unexplored thus far. It might also be possible to include segmentation into a longitudinal registration and bias correction technique, combining the approach of CLASSIC [75] with that of SPM5 [27], though further work would be required to explore the feasibility and desirability of such a complex model.

Finally, the question of whether sets of more than two images can be treated in any better way than via separate pairwise differential bias correction should be addressed. Note that Learned-Miller and Jain [47] developed a method that can use multiple subject images (not necessarily serial repeat scans), but this method requires of the order of 20 or more images, which would very rarely be applicable to single-subject longitudinal correction. Their technique is also more focussed on statistically removing common components of bias from potentially quite different images, rather than more directly eliminating the differential bias between approximately similar images. Combined multi-image DBC and registration would be a challenging task, since registration techniques for more than two images are themselves a topic of current research [76].

5.2.8 Future Experiments

To begin with, the experiment discussed here should additionally evaluate the use of bias correction after a complete multi-level non-rigid registration step without interspersed bias correction, to prove that the interleaved method is beneficial. It would also be interesting to test the effect of the combined algorithm on the simulated atrophy data without simulated bias — does the method falsely detect differential bias? And how does the registration performance compare without bias correction on this data without differential bias. Beyond these investigations, there is great scope for further experimental research.

The pilot study presented above used only a very small number of images, simply to see whether the observed behaviour showed any apparently consistent trends. Future work should involve a much greater number of simulated atrophy images, mimicking a range of disease severities and perhaps healthy controls. Severe atrophy, and perhaps the presence of additional artefacts such as simulated motion or pulsatile flow [77], would present interesting challenges to the combined DBC and registration approach. It will also

[10] Further details can be found at `http://www.fmrib.ox.ac.uk/fsl/fnirt/index.html`.

be important to test the algorithm on real data, though indirect methods of performance quantification will then be required. Lewis and Fox [16] reported two such quantitative tests, using the alteration in results from brain BSI [15] to reflect on bias correction performance. In the first, same-day scan pairs were used to investigate whether DBC induced false changes in brain volume. In the second, the effect of DBC on brain BSI was investigated in one group of subjects with negligible visually-assessed differential bias, and one group with more significant differential bias. They found that DBC reduced the amount of atrophy found in both groups, suggesting that some genuine atrophy was being removed by the process. There is reason to hope that combined longitudinal registration and bias correction will be more successful here, since the final bias correction step on the aligned images will have greatly reduced opportunity to 'correct' for intensity differences due to the atrophic mis-alignment of the original image pair. The experiments of Lewis and Fox [16] should therefore be replicated with the new algorithm. Similar data could be used to optimise filtering and masking options discussed above.

Both the simulated atrophy images and the real scans used in the original DBC paper were acquired at 1.5T, an interesting and important avenue for further investigation is the correction of higher field image sets. The Neurogrid project [78] has images of the same subjects acquired at both 1.5 and 3T, which could allow a useful comparison of performance; broadly speaking, the better the algorithm, the more similar the two field strength's images should appear (though other changes in contrast and SNR will of course mean that identical images cannot be the goal). In Huntington's Disease [79], there is great interest in the caudate — a structure whose main source of confounding artefact in MRI is intensity non-uniformity. The HADNI project within the TrackHD venture[11] could be a very important opportunity for serial registration-based bias correction to contribute to clinical research. The data includes 3T images. It may also be a valuable contribution to the field to investigate more realistic simulation of intensity non-uniformity at high field, perhaps including a simple brain segmentation to model object-dependent RF penetration and/or standing wave effects, if feasible. Lewis and Fox [16] did not compare the use of their differential bias correction algorithm to the separate application of any single-image bias correction techniques. Such comparison, with a popular and respected algorithm such as N3 [28], seems essential to motivate the use of DBC. It would also be interesting (if possible) to compare the method to the implementation of template-based bias correction by Studholme et al. [48], which appeared to be particularly successful.

5.2.9 Conclusion

The important problem of correcting differential bias in serial MR images. has been introduced, and background information on the sources of the bias field, and on existing bias correction methods has been given. An existing DBC method has been extended and integrated within a multi-level non-rigid registration algorithm. Evaluation on simulated atrophy data showed improvements to both bias-correction and registration from their combination. Potential ideas for more novel developments have been explored. In closing,

[11] `http://www.track-hd.net`

it is important to reiterate that this work is of great relevance to other research in this thesis. More accurate longitudinal registration and differential bias correction would directly impact on deformation- or tensor-based morphometry, as well as longitudinal methods of voxel-based morphometry. The information derived from the deformation fields should be more reliable and valid, whether used directly (for example in multivariate analysis of the strain tensor) or indirectly (such as through Jacobian-modulation of grey-matter segments).

Bibliography

[1] R. P. Woods, "Characterizing volume and surface deformations in an atlas framework: theory, applications, and implementation." *Neuroimage*, vol. 18, no. 3, pp. 769–788, Mar. 2003. ˆ303, 304, 305, 306, 307, 308, 311

[2] V. Arsigny, "Processing data in Lie groups: An algebraic approach. application to non-linear registration and diffusion tensor MRI," Ph.D. dissertation, École Polytechnique, Paris, 2006. ˆ304, 305, 306, 307, 308, 311

[3] P. T. Fletcher, L. Conglin, S. M. Pizer, and S. Joshi, "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE Trans. Med. Imag.*, vol. 23, no. 8, pp. 995–1005, Aug. 2004. ˆ305

[4] N. Lepore, C. Brun, Y. Y. Chou, M. C. Chiang, R. A. Dutton, K. M. Hayashi, E. Luders, O. L. Lopez, H. J. Aizenstein, A. W. Toga, J. T. Becker, and P. M. Thompson, "Generalized tensor-based morphometry of HIV/AIDS using multivariate statistics on deformation tensors." *IEEE Trans. Med. Imag.*, vol. 27, no. 1, pp. 129–141, Jan. 2008. ˆ306

[5] P. Fillard, V. Arsigny, X. Pennec, K. M. Hayashi, P. M. Thompson, and N. Ayache, "Measuring brain variability by extrapolating sparse tensor fields measured on sulcal lines." *Neuroimage*, vol. 34, no. 2, pp. 639–650, Jan. 2007. ˆ307

[6] M. Moakher, "Means and averaging in the group of rotations," *SIAM Journal on matrix analysis and applications*, vol. 24, no. 1, pp. 1–16, 2002. ˆ307, 308, 310

[7] M. Bossa, M. Hernandez, and S. Olmos, "Contributions to 3D diffeomorphic atlas estimation: application to brain images." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 10, no. Pt 1, 2007, pp. 667–674. [Online]. Available: http://www.springerlink.com/content/e787205265muu244/ ˆ308

[8] V. Arsigny, O. Commowick, X. Pennec, and N. Ayache, "A log-Euclidean framework for statistics on diffeomorphisms." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 9, no. Pt 1, 2006, pp. 924–931. [Online]. Available: http://www.springerlink.com/content/607206763v078397/ ˆ308

[9] M. Alexa, "Linear combination of transformations," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 380–387, 2002. ˆ309

[10] C. Bloom, J. Blow, and C. Muratori, "Errors and omissions in Marc Alexa's "linear combinations of transformations"," 2004, unfinished draft. [Online]. Available: http://www.cbloom.com/3d/techdocs/lcot_errors.pdf ˆ309

[11] F. L. Bookstein, "Landmark methods for forms without landmarks: morphometrics of group differences in outline shape." *Med Image Anal*, vol. 1, no. 3, pp. 225–243, Apr. 1997. ˆ309

[12] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C++*. Cambridge University Press, 2002. ˆ310

[13] B. Whitcher, J. J. Wisco, N. Hadjikhani, and D. S. Tuch, "Statistical group comparison of diffusion tensors via multivariate hypothesis testing." *Magn Reson Med*, vol. 57, no. 6, pp. 1065–1074, Jun. 2007. ˆ310

[14] A. Simmons, P. S. Tofts, G. J. Barker, and S. R. Arridge, "Sources of intensity nonuniformity in spin echo images at 1.5 T." *Magn Reson Med*, vol. 32, no. 1, pp. 121–128, Jul. 1994. ˆ312, 313, 314, 315, 318

[15] P. A. Freeborough and N. C. Fox, "The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI." *IEEE Trans. Med. Imag.*, vol. 16, no. 5, pp. 623–629, Oct. 1997. ˆ312, 329

[16] E. B. Lewis and N. C. Fox, "Correction of differential intensity inhomogeneity in longitudinal MR images." *Neuroimage*, vol. 23, no. 1, pp. 75–83, Sep. 2004. ˆ312, 313, 317, 318, 319, 320, 325, 326, 328, 329

[17] S. M. Smith, Y. Zhang, M. Jenkinson, J. Chen, P. M. Matthews, A. Federico, and N. D. Stefano, "Accurate, robust, and automated longitudinal and cross-sectional brain change analysis." *Neuroimage*, vol. 17, no. 1, pp. 479–489, Sep. 2002. ˆ312

[18] C. Studholme, C. Drapaca, B. Iordanova, and V. Cardenas, "Deformation-based mapping of volume change from serial brain MRI in the presence of local tissue contrast change." *IEEE Trans. Med. Imag.*, vol. 25, no. 5, pp. 626–639, May 2006. ˆ312, 319, 327

[19] D. L. Pham and J. L. Prince, "Adaptive fuzzy segmentation of magnetic resonance images." *IEEE Trans. Med. Imag.*, vol. 18, no. 9, pp. 737–752, Sep. 1999. ˆ312

[20] U. Vovk, F. Pernus, and B. Likar, "A review of methods for correction of intensity inhomogeneity in MRI," *IEEE Trans. Med. Imag.*, vol. 26, no. 3, pp. 405–421, Mar. 2007. ˆ312, 316, 318, 326, 328

[21] D. A. Wicks, G. J. Barker, and P. S. Tofts, "Correction of intensity nonuniformity in MR images of any orientation." *Magn Reson Imaging*, vol. 11, no. 2, pp. 183–196, 1993. ˆ312

[22] S.-H. Lai and M. Fang, "A dual image approach for bias field correction in magnetic resonance imaging." *Magn Reson Imaging*, vol. 21, no. 2, pp. 121–125, Feb. 2003. ˆ312, 327

[23] D. L. Thomas, E. D. Vita, R. Deichmann, R. Turner, and R. J. Ordidge, "3D MDEFT imaging of the human brain at 4.7 T with reduced sensitivity to radiofrequency inhomogeneity." *Magn Reson Med*, vol. 53, no. 6, pp. 1452–1458, Jun. 2005. ˆ312, 313, 314, 315

[24] B. Brinkmann, A. Manduca, and R. Robb, "Optimized homomorphic unsharp masking for MR grayscale inhomogeneity correction," *IEEE Trans. Med. Imag.*, vol. 17, no. 2, pp. 161–171, Apr. 1998. ˆ312, 315, 317, 318, 325

[25] B. Dawant, A. Zijdenbos, and R. Margolin, "Correction of intensity variations in MR images for computer-aided tissue classification," *IEEE Trans. Med. Imag.*, vol. 12, no. 4, pp. 770–781, 1993. ˆ312, 326

[26] W. M. Wells, III, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz, "Adaptive segmentation of MRI data," *IEEE Trans. Med. Imag.*, vol. 15, pp. 429–442, 1996. ˆ312

[27] J. Ashburner and K. J. Friston, "Unified segmentation." *Neuroimage*, vol. 26, no. 3, pp. 839–851, Jul. 2005. ˆ312, 316, 327, 328

[28] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data." *IEEE Trans. Med. Imag.*, vol. 17, no. 1, pp. 87–97, Feb. 1998. ˆ312, 313, 314, 315, 316, 318, 319, 321, 326, 329

[29] J. B. Arnold, J. S. Liow, K. A. Schaper, J. J. Stern, J. G. Sled, D. W. Shattuck, A. J. Worth, M. S. Cohen, R. M. Leahy, J. C. Mazziotta, and D. A. Rottenberg, "Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects." *Neuroimage*, vol. 13, no. 5, pp. 931–943, May 2001. ˆ312, 315, 316, 318, 319

[30] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy, "Magnetic resonance image tissue classification using a partial volume model." *Neuroimage*, vol. 13, no. 5, pp. 856–876, May 2001. ˆ312, 316, 319, 326

[31] J. V. Hajnal, N. Saeed, E. J. Soar, A. Oatridge, I. R. Young, and G. M. Bydder, "A registration and interpolation procedure for subvoxel matching of serially acquired MR images." *J Comput Assist Tomogr*, vol. 19, no. 2, pp. 289–296, 1995. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/7890857 ˆ312

[32] G. E. Christensen, R. D. Rabbitt, and M. I. Miller, "Deformable templates using large deformation kinematics," *IEEE Trans. Image Process.*, vol. 5, no. 10, pp. 1435–1447, Oct. 1996. ˆ312, 327

[33] P. A. Freeborough and N. C. Fox, "Modeling brain deformations in Alzheimer disease by fluid registration of serial 3D MR images." *J Comput Assist*

Tomogr, vol. 22, no. 5, pp. 838–843, 1998. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9754126 ˆ312

[34] J. G. Sled and G. B. Pike, "Standing-wave and RF penetration artifacts caused by elliptic geometry: an electrodynamic analysis of MRI." *IEEE Trans. Med. Imag.*, vol. 17, no. 4, pp. 653–662, Aug. 1998. ˆ313, 314, 315

[35] M. Alecci, C. M. Collins, M. B. Smith, and P. Jezzard, "Radio frequency magnetic field mapping of a 3 Tesla birdcage coil: experimental and theoretical dependence on sample properties." *Magn Reson Med*, vol. 46, no. 2, pp. 379–385, Aug. 2001. ˆ313, 314

[36] F. B. Mohamed, S. Vinitski, S. H. Faro, H. V. Ortega, and S. Enochs, "A simple method to improve image nonuniformity of brain MR images at the edges of a head coil." *J Comput Assist Tomogr*, vol. 23, no. 6, pp. 1008–1012, 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10589586 ˆ313

[37] D. Wang and D. M. Doddrell, "Method for a detailed measurement of image intensity nonuniformity in magnetic resonance imaging." *Med Phys*, vol. 32, no. 4, pp. 952–960, Apr. 2005. ˆ313

[38] M. S. Cohen, R. M. DuBois, and M. M. Zeineh, "Rapid and effective correction of RF inhomogeneity for high field magnetic resonance imaging." *Hum Brain Mapp*, vol. 10, no. 4, pp. 204–211, Aug. 2000. ˆ314, 318

[39] E. De Vita, D. L. Thomas, S. Roberts, H. G. Parkes, R. Turner, P. Kinchesh, K. Shmueli, T. A. Yousry, and R. J. Ordidge, "High resolution MRI of the brain at 4.7 Tesla using fast spin echo imaging." *Br J Radiol*, vol. 76, no. 909, pp. 631–637, Sep. 2003. ˆ314

[40] S. Balac and L. Chupin, "Fast approximate solution of Bloch equation for simulation of RF artifacts in magnetic resonance imaging," Institut Camille Jordan, Tech. Rep., 2007. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00138491/fr/ ˆ314

[41] H. Benoit-Cattin, G. Collewet, B. Belaroussi, H. Saint-Jalmes, and C. Odet, "The SIMRI project: a versatile and interactive MRI simulator." *J Magn Reson*, vol. 173, no. 1, pp. 97–115, Mar. 2005. ˆ314

[42] C. Cocosco, V. Kollokian, R. Kwan, and A. Evans, "Brainweb: Online interface to a 3D MRI simulated brain database," *NeuroImage*, vol. 5, no. 4, p. S425, 1997. [Online]. Available: http://www.bic.mni.mcgill.ca/users/crisco/HBM97_abs/HBM97_abs.html ˆ315

[43] J. D. Gispert, S. Reig, J. Pascau, J. J. Vaquero, P. García-Barreno, and M. Desco, "Method for bias field correction of brain T1-weighted magnetic resonance images minimizing segmentation error." *Hum Brain Mapp*, vol. 22, no. 2, pp. 133–144, Jun. 2004. ˆ315

[44] A. D. Leow, A. D. Klunder, C. R. Jack, A. W. Toga, A. M. Dale, M. A. Bernstein, P. J. Britson, J. L. Gunter, C. P. Ward, J. L. Whitwell, B. J. Borowski, A. S. Fleisher, N. C. Fox, D. Harvey, J. Kornak, N. Schuff, C. Studholme, G. E. Alexander, M. W. Weiner, and P. M. Thompson, "Longitudinal stability of MRI for mapping brain change using tensor-based morphometry." *Neuroimage*, vol. 31, no. 2, pp. 627–640, Jun. 2006, A.D.N.I. Preparatory Phase Study. ˆ316

[45] P. A. Bromiley and N. A. Thacker, "A model-independent, multi-image approach to MR inhomogeneity correction," in *Medical Image Understanding and Analysis*, 2007. ˆ316

[46] E. A. Vokurka, N. A. Thacker, and A. Jackson, "A fast model independent method for automatic correction of intensity nonuniformity in MRI data." *J Magn Reson Imaging*, vol. 10, no. 4, pp. 550–562, Oct. 1999. ˆ316

[47] E. G. Learned-Miller and V. Jain, "Many heads are better than one: jointly removing bias from multiple MRIs using nonparametric maximum likelihood." in *Inf. Process. Med. Imag.*, vol. 19, 2005, pp. 615–626. [Online]. Available: http://www.springerlink.com/content/709hh1ny9ydfeh44/ ˆ316, 328

[48] C. Studholme, V. Cardenas, E. Song, F. Ezekiel, A. Maudsley, and M. Weiner, "Accurate template-based correction of brain MRI intensity distortion with application to dementia and aging," *IEEE Trans. Med. Imag.*, vol. 23, no. 1, pp. 99–110, Jan. 2004. ˆ317, 318, 319, 324, 325, 326, 327, 329

[49] L. Axel, J. Costantini, and J. Listerud, "Intensity correction in surface-coil MR imaging." *AJR Am J Roentgenol*, vol. 148, no. 2, pp. 418–420, Feb. 1987. ˆ318, 325

[50] B. Johnston, M. Atkins, B. Mackiewich, and M. Anderson, "Segmentation of multiple sclerosis lesions in intensity corrected multispectral MRI," *IEEE Trans. Med. Imag.*, vol. 15, no. 2, pp. 154–169, Apr. 1996. ˆ318, 325

[51] H. Gudbjartsson and S. Patz, "The Rician distribution of noisy MRI data." *Magn Reson Med*, vol. 34, no. 6, pp. 910–914, Dec. 1995. ˆ319

[52] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images." *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999. ˆ319, 320

[53] J. A. Schnabel, D. Rueckert, M. Quist, J. M. Blackall, A. D. Castellano-Smith, T. Hartkens, G. P. Penney, W. A. Hall, H. Liu, C. L. Truwit, F. A. Gerritsen, D. L. G. Hill, and D. J. Hawkes, "A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 2208, Jan. 2001, pp. 573–. [Online]. Available: http://www.springerlink.com/content/0q3bfeunq4avrwdf/ ˆ319, 324, 326

[54] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "Overlap invariant entropy measure of 3D medical image alignment," *Pattern Recognition*, vol. 32, no. 1, pp. 71–86, Jan. 1999. ˆ320

[55] R. G. Boyes, D. Rueckert, P. Aljabar, J. Whitwell, J. M. Schott, D. L. G. Hill, and N. C. Fox, "Cerebral atrophy measurements using Jacobian integration: comparison with the boundary shift integral." *Neuroimage*, vol. 32, no. 1, pp. 159–169, Aug. 2006. ˆ320

[56] J. Ashburner, J. L. Andersson, and K. J. Friston, "High-dimensional image registration using symmetric priors." *Neuroimage*, vol. 9, no. 6 Pt 1, pp. 619–628, Jun. 1999. ˆ320

[57] W. R. Crum, O. Camara, and D. J. Hawkes, "Methods for inverting dense displacement fields: evaluation in brain image registration." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 10, no. Pt 1, 2007, pp. 900–907. [Online]. Available: http://www.springerlink.com/content/132m22p1w68p1350/ ˆ320

[58] S. Lee, G. Wolberg, and S. Shin, "Scattered data interpolation with multilevel B-splines," *IEEE Trans. Vis. Comput. Graphics*, vol. 3, no. 3, pp. 228–244, 1997. ˆ320

[59] O. Camara, M. Schweiger, R. Scahill, W. Crum, B. Sneller, J. Schnabel, G. Ridgway, D. Cash, D. Hill, and N. Fox, "Phenomenological model of diffuse global and regional atrophy using finite-element methods," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1417–1430, Nov. 2006. ˆ320

[60] W. R. Crum, L. D. Griffin, D. L. G. Hill, and D. J. Hawkes, "Zen and the art of medical image registration: correspondence, homology, and quality." *Neuroimage*, vol. 20, no. 3, pp. 1425–1437, Nov. 2003. ˆ324

[61] Z. Xue, D. Shen, and C. Davatzikos, "Statistical representation of high-dimensional deformation fields with application to statistically constrained 3D warping." *Med Image Anal*, vol. 10, no. 5, pp. 740–751, Oct. 2006. ˆ324

[62] C. Tanner, J. Schnabel, A. Degenhard, A. Castellano-Smith, C. Hayes, M. Leach, D. Hose, D. Hill, and D. Hawkes, "Validation of volume-preserving non-rigid registration: Application to contrast-enhanced MR-mammography," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 2488. Springer, 2002, pp. 307–314. [Online]. Available: http://www.springerlink.com/content/01f403a1uwe825ve/ ˆ324

[63] T. Rohlfing, C. R. Maurer, D. A. Bluemke, and M. A. Jacobs, "Volume-preserving nonrigid registration of MR breast images using free-form deformation with an incompressibility constraint." *IEEE Trans. Med. Imag.*, vol. 22, no. 6, pp. 730–741, Jun. 2003. ˆ324

[64] J. Bednar and T. Watt, "Alpha-trimmed means and their relationship to median filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 1, pp. 145–153, 1984. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1164279 ˆ326

[65] A. Restrepo and A. Bovik, "Adaptive trimmed mean filters for image restoration," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 8, pp. 1326–1337, 1988. ˆ326

[66] S. Kassam and H. Poor, "Robust techniques for signal processing: A survey," *Proceedings of the IEEE*, vol. 73, no. 3, pp. 433–481, 1985. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1457435 ˆ326

[67] R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing.* Academic Press, 2005. ˆ326

[68] P. Holland and R. Welsch, "Robust regression using iteratively reweighted least-squares," *Communications in Statistics-Theory and Methods*, vol. 6, no. 9, pp. 813–827, 1977. ˆ327

[69] R. Subbarao and P. Meer, "Beyond RANSAC: User independent robust regression," in *Computer Vision and Pattern Recognition Workshop, 2006 Conference on*, Jun. 2006, pp. 101–101. ˆ327

[70] W. R. Crum, C. Tanner, and D. J. Hawkes, "Anisotropic multi-scale fluid registration: evaluation in magnetic resonance breast imaging." *Phys Med Biol*, vol. 50, no. 21, pp. 5153–5174, Nov. 2005. ˆ327

[71] Z. Xue, D. Shen, and C. Davatzikos, "Determining correspondence in 3-D MR brain images using attribute vectors as morphological signatures of voxels." *IEEE Trans. Med. Imag.*, vol. 23, no. 10, pp. 1276–1291, Oct. 2004. ˆ327

[72] P. Cachier, J.-F. Mangin, X. Pennec, D. Rivière, D. Papadopoulos-Orfanos, J. Régis, and N. Ayache, "Multisubject non-rigid registration of brain MRI using intensity and geometric features," in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, ser. LNCS, vol. 2208, Jan. 2001, pp. 734–. [Online]. Available: http://www.springerlink.com/content/jyjb1d0dwt8cy03y/ ˆ327

[73] D. Loeckx, P. Slagmolen, F. Maes, D. Vandermeulen, and P. Suetens, "Nonrigid image registration using conditional mutual information." in *Inf. Process. Med. Imag.*, vol. 20, 2007, pp. 725–737. [Online]. Available: http://www.springerlink.com/content/e61n05633j6631k5/ ˆ327

[74] Z. F. Knops, J. B. A. Maintz, M. A. Viergever, and J. P. W. Pluim, "Normalized mutual information based registration using k-means clustering and shading correction." *Med Image Anal*, vol. 10, no. 3, pp. 432–439, Jun. 2006. ˆ328

[75] Z. Xue, D. Shen, and C. Davatzikos, "CLASSIC: consistent longitudinal alignment and segmentation for serial image computing." *Neuroimage*, vol. 30, no. 2, pp. 388–399, Apr. 2006. ˆ328

[76] J. Ashburner, "A fast diffeomorphic image registration algorithm." *Neuroimage*, vol. 38, no. 1, pp. 95–113, Oct. 2007. ˆ328

[77] O. Camara-Rey, B. I. Sneller, G. R. Ridgway, E. Garde, N. C. Fox, and D. L. G. Hill, "Simulation of acquisition artefacts in MR scans: effects on automatic measures of brain atrophy." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 9, no. Pt 1, 2006, pp. 272–280. [Online]. Available: http://www.springerlink.com/content/l2x6220286r78578/ ˆ328

[78] J. Geddes, S. Lloyd, A. Simpson, M. Rossor, N. Fox, D. Hill, J. Hajnal, S. Lawrie, A. Mclntosh, E. Johnstone, J. Wardlaw, D. Perry, R. Procter, P. Bath, and E. Bullmore, "NeuroGrid: using grid technology to advance neuroscience," in *18th IEEE Symposium on Computer-Based Medical Systems*, Jun. 2005, pp. 570–572. ˆ329

[79] B. Harper, "Huntington disease." *J R Soc Med*, vol. 98, no. 12, p. 550, Dec. 2005. ˆ329

Chapter 6

Conclusion

In this thesis, an attempt has been made to address a particularly broad range of issues, even within a field that is well-known for requiring inter-disciplinary approaches and varied expertise. New permutation-testing strategies for general linear models have been developed and thoroughly evaluated, and theoretical connections between existing methods strengthened. Practical methodological developments have been made with regard to the popular technique of voxel-based morphometry, and the resulting improvements have been demonstrated on real and simulated data. Generalised tensor-based morphometry methods and theory have been advanced, particularly with regard to the number and type of measurements available for testing different aspects of morphometric change. We have begun to address the problem of differential bias, considered to be of great importance to the subsequent analysis of longitudinal MRI data. Despite the attempt at breadth, several parts of the work have been performed in great depth and with close attention to detail. We have detected some flaws in published methods of tensor reorientation, and explained carefully how they may be rectified. We have also endeavoured, especially during the first chapter, but also throughout the thesis, to keep in mind the possible clinical applications of the work. This is most obvious in the chapter on voxel-based morphometry, where the masking section directly studies and solves a problem identified in VBM of neurodegeneration, but further examples can be found in the attempts to visualise and interpret complex findings from generalised TBM.

Another goal of this thesis has been to open up avenues for further exploration. Every major segment of work has been accompanied with suggestions or more detailed plans for future research. Several of the ideas proposed seem to have major potential, including the suggestion that a particular permutation-testing strategy could allow much more rigorous analysis of unnormalised 'contrast' maps than is currently available in the literature. Another particularly notable new idea is the potential for Riemannian Cramér testing of strain (or diffusion) tensors and their principal eigenvectors.

The original aim of the work was to develop methods of voxel-wise statistical analysis for application to data derived from longitudinal structural MRI. In fact, most of the significant developments in the thesis are more general than this, for example with application to cross-sectional morphometry or even functional imaging. Much of the work on tensor-based morphometry is of immediate relevance to diffusion tensor image analy-

sis. The permutation-testing methods are of very wide applicability, though, somewhat ironically, are not particularly well-suited to longitudinal data, which typically requires iteratively computed mixed-effects models, with which the additional computational demands of resampling methods are not currently compatible. We hope to begin to address this challenging task in further work, though we note that a simple approach using (longitudinal) summary statistics is immediately available within the permutation-testing framework proposed here.

We believe that serial MR imaging still has much more to offer in the analysis of neurodegenerative diseases, and that both the image processing and the statistical modelling must be further tailored to the longitudinal nature of the data in order to draw the most value from such studies. It is hoped that this thesis contributes both to this application and to its supporting methodology.

6.1 Summary of contributions by chapter

6.1.1 Introduction

- Concise summary of clinical challenges.

6.1.2 Permutation Testing

- Theory of permutation testing methods for arbitrary linear models, including multivariate data. FSL's randomise and SnPM only handle univariate data, and have previously been incorrect for arbitrary linear models.
- Thorough Monte Carlo evaluation of different permutation strategies, including novel formulations.
- Implementation of permutation testing. General code, allowing non-standard statistics such as Cramér, and the use of searchlight neighborhoods. Time- and memory-efficient parallel code (allows higher-resolution data than randomise/SnPM).

6.1.3 Voxel-based Morphometry

- Discovery of potential problem with standard SPM mask-creation strategy, and development of two original solutions.
- Implementation of new methods for combined longitudinal registration and VBM preprocessing, one is entirely novel, the rest are newly implemented to work with SPM5's unified segmentation.
- Investigation of the relative performance of the VBM methods using a novel approach to the generation of gold standard results, employing an atrophy simulation model.
- An original summary of important aspects related to reporting VBM studies (joint work with others at DRC).

6.1.4 Multivariate Morphometry

- Theoretical concepts. Synthesis of Hencky tensor (from solid mechanics) and log-Euclidean analysis of strain tensor (mathematically motivated). Clarification and extension of methods of tensor reorientation.
- Presentation of FWE-corrected results on multivariate tensor-based morphometry. Other work has used less strict FDR-correction or no correction at all for multiple comparisons.
- First use of Cramér test and Watson test for morphometry data.
- Use of searchlight method on structural imaging data, plus extension to FWE correction, and use of scale-pyramid approach.

6.1.5 Further Developments

- Explanation of cutting edge theory regarding the semi-Riemannian or affine-invariant mean of Jacobian tensors.
- Practical suggestion for implementation of above Jacobian tensor analysis.
- Differential bias correction. Iterative registration and differential bias correction. Novel proposal to extend existing work. Basic implementation, using multi-level Free-Form Deformation registration. Original ideas for further work, in particular the use of novel trimmed-mean filtering for DBC.

6.2 Coauthored Publications

6.2.1 Journal

- An investigation into a potential problem with the standard method for defining the set of voxels analysed in VBM, with suggestions of an improved technique [1].
- Guidelines for reporting VBM studies, discussing some of the more subtle issues and highlighting some potential pitfalls [2].
- Fronto-temporal lobar degeneration investigated using vertex-wise statistical analysis of cortical thickness [3].
- A VBM study of Huntington's disease (HD), looking for relationships between regional brain volumes and a measure of the genetic severity of the disease[4].
- A VBM study of facial emotion recognition in HD[5].
- Evaluation of brain boundary shift integral and Jacobian integration atrophy measurement techniques using realistic simulated atrophy [6].
- Post-mortem imaging, including non-invasive thermometry using the apparent diffusion coefficient, and temperature-adjusted FLAIR imaging [7].

- The simulation of serial MRI exhibiting AD-like atrophy, using a finite element method to deform the baseline image [8].
- A fast implementation of a B-spline free-form deformation registration algorithm on multi-core graphics hardware [9].

6.2.2 Refereed conference

- A formulation of the analytical gradient of the normalised mutual information registration criterion, suitable for parallel implementation on a graphics card [10].
- Evaluation of atropy measurement using simulated atrophy [11].
- Evaluation of atropy measurement using simulated atrophy (MIUA Best Paper Award) [12].
- Implementation and comparison of four methods of longitudinal VBM evaluated using simulated atrophy [13].
- Simulation of motion and pulsation artefacts in MRI, and investigation of their effect on automatic atrophy measurement [14].

6.2.3 Conference abstracts

- Posterior cortical atrophy (the 'visual AD') studied with VBM and vertex-wise analysis of cortical thickness [15].
- Exploration of common morphometric 'nuisance' covariates (age, gender and head size) in a healthy control population [16].
- Multivariate generalised TBM and searchlight-based morphometry on a longitudinal AD cohort (NIH Travel Award) [17].
- Monte-Carlo evaluation of size and power of different strategies for permutation-testing with general linear models [18].
- Comparison of different approaches for longitudinal VBM, and TBM-like voxel-compression mapping on real AD data [19].
- A VBM study of the morphometric effect of a candidate anti-amyloid drug, comparing treatment- and placebo-group AD patients [20].

Bibliography

[1] G. R. Ridgway, R. Omar, S. Ourselin, D. L. G. Hill, J. D. Warren, and N. C. Fox, "Issues with threshold masking in voxel-based morphometry of atrophied brains," *NeuroImage*, vol. 44, no. 1, pp. 99–111, Jan. 2009. ˆ340

[2] G. R. Ridgway, S. M. D. Henley, J. D. Rohrer, R. I. Scahill, J. D. Warren, and N. C. Fox, "Ten simple rules for reporting voxel-based morphometry studies." *Neuroimage*, vol. 40, no. 4, pp. 1429–1435, May 2008. ˆ340

[3] J. D. Rohrer, J. D. Warren, M. Modat, G. R. Ridgway, A. Douiri, M. N. Rossor, S. Ourselin, and N. C. Fox, "Patterns of cortical thinning in the language variants of frontotemporal lobar degeneration." *Neurology*, vol. 72, no. 18, pp. 1562–1569, May 2009. ˆ340

[4] S. M. D. Henley, E. J. Wild, N. Z. Hobbs, R. I. Scahill, G. R. Ridgway, D. G. MacManus, R. A. Barker, N. C. Fox, and S. J. Tabrizi, "Relationship between CAG repeat length and brain volume in premanifest and early Huntington's disease." *Journal of Neurology*, vol. 256, pp. 203–212, 2009. ˆ340

[5] S. M. D. Henley, E. J. Wild, N. Z. Hobbs, J. D. Warren, C. Frost, R. I. Scahill, G. R. Ridgway, D. G. MacManus, R. A. Barker, N. C. Fox, and S. J. Tabrizi, "Defective emotion recognition in early hd is neuropsychologically and anatomically generic." *Neuropsychologia*, vol. 46, no. 8, pp. 2152–2160, 2008. ˆ340

[6] O. Camara, J. A. Schnabel, G. R. Ridgway, W. R. Crum, A. Douiri, R. I. Scahill, D. L. G. Hill, and N. C. Fox, "Accuracy assessment of global and local atrophy measurement techniques with realistic simulated longitudinal Alzheimer's disease images." *Neuroimage*, vol. 42, no. 2, pp. 696–709, Aug. 2008. ˆ340

[7] P. S. Tofts, J. S. Jackson, D. J. Tozer, M. Cercignani, G. Keir, D. G. Macmanus, G. R. Ridgway, B. H. Ridha, K. Schmierer, D. Siddique, J. S. Thornton, S. J. Wroe, and N. C. Fox, "Imaging cadavers: Cold FLAIR and noninvasive brain thermometry using CSF diffusion." *Magn Reson Med*, vol. 59, no. 1, pp. 190–195, Jan. 2008. ˆ340

[8] O. Camara, M. Schweiger, R. Scahill, W. Crum, B. Sneller, J. Schnabel, G. Ridgway, D. Cash, D. Hill, and N. Fox, "Phenomenological model of diffuse global and regional atrophy using finite-element methods," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1417–1430, Nov. 2006. ˆ341

[9] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin, "Fast free-form deformation using graphics processing units." *Comput Methods Programs Biomed*, 2009, in press. [Online]. Available: http://eprints.ucl.ac.uk/17431 ˆ341

[10] M. Modat, G. Ridgway, Z. Taylor, D. Hawkes, N. Fox, and S. Ourselin, "A parallel-friendly normalised mutual information gradient for free-form deformation," in *SPIE Medical Imaging*, 2009. ˆ341

[11] O. Camara, R. I. Scahill, J. A. Schnabel, W. R. Crum, G. R. Ridgway, D. L. G. Hill, and N. C. Fox, "Accuracy assessment of global and local atrophy measurement techniques with realistic simulated longitudinal data." in *Int. Conf. Med. Image*

Comput. Comput. Assisted Intervention, vol. 10, no. Pt 2, 2007, pp. 785–792. [Online]. Available: http://www.springerlink.com/content/w18323021vhnx430/ ˆ341

[12] O. Camara, R. Scahill, W. Crum, J. Schnabel, G. Ridgway, D. Hill, and N. Fox, "Evaluation of local and global atrophy measurement techniques with simulated Alzheimer's disease images," in *Medical Image Understanding and Analysis*, 2007, pp. 16–20. [Online]. Available: http://users.aber.ac.uk/rrz/miua2007_proceedings.pdf ˆ341

[13] G. Ridgway, O. Camara, R. Scahill, W. Crum, B. Whitcher, N. Fox, and D. Hill, "Longitudinal voxel-based morphometry with unified segmenation: Evaluation on simulated Alzheimer's disease," in *Medical Image Understanding and Analysis*, 2007, pp. 201–205. [Online]. Available: http://users.aber.ac.uk/rrz/miua2007_proceedings.pdf ˆ341

[14] O. Camara-Rey, B. I. Sneller, G. R. Ridgway, E. Garde, N. C. Fox, and D. L. G. Hill, "Simulation of acquisition artefacts in MR scans: effects on automatic measures of brain atrophy." in *Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, vol. 9, no. Pt 1, 2006, pp. 272–280. [Online]. Available: http://www.springerlink.com/content/l2x6220286r78578/ ˆ341

[15] M. Lehmann, S. J. Crutch, G. R. Ridgway, B. H. Ridha, J. Barnes, E. K. Warrington, M. N. Rossor, and N. C. Fox, "Cortical thickness and voxel-based morphometry in posterior cortical atrophy and typical alzheimer's disease." *Neurobiol Aging*, Sep. 2009. ˆ341

[16] J. Barnes, S. Henley, M. Lehmann, N. Hobbs, R. Scahill, M. Clarkson, G. Ridgway, D. MacManus, S. Ourselin, and N. Fox, "Head size, age and gender adjustment in MRI studies: A necessary nuisance?" *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, vol. 5, no. 4S, pp. 102–103, 2009. ˆ341

[17] G. R. Ridgway, B. Whitcher, D. L. G. Hill, and N. C. Fox, "Longitudinal multivariate tensor- and searchlight-based morphometry using permutation testing," in *14th annual meeting of the Organization for Human Brain Mapping*, 2008. ˆ341

[18] T. Nichols, G. Ridgway, M. Webster, and S. Smith, "GLM permutation - nonparametric inference for arbitrary general linear models," in *14th annual meeting of the Organization for Human Brain Mapping*, 2008. ˆ341

[19] G. R. Ridgway, R. I. Scahill, D. L. G. Hill, and N. C. Fox, "A comparison of voxel compression mapping and longitudinal voxel-based morphometry," in *13th annual meeting of the Organization for Human Brain Mapping*, 2007. [Online]. Available: http://eprints.ucl.ac.uk/13846/ ˆ341

[20] R. Scahill, G. Ridgway, R. Black, M. Grundman, D. Hill, and N. Fox, "IC-104-01 Regional distribution of grey matter changes in Abeta (AN1792) immunized patients

with AD: A voxel-based morphometry analysis," in *ICAD*, vol. 2, no. 3S. Madrid: Elsevier, 2006, pp. 654–654. ˆ341

Appendix A

Mathematics for the linear model

This appendix presents some mathematical background, together with key results and derivations, where relevant to other topics considered in the thesis. In particular, it focuses on the general linear model, and the supporting concepts of subspaces and projection matrices. It includes material on the multivariate general linear model which is less common in the neuroimaging literature. An attempt has been made to provide more thorough and/or more logically motivated derivations of results that are frequently stated in isolation, as well as a more complete presentation of reformulated variants and extensions. Section A.4.8 includes a novel proof of a simplified result for the partitioned form of a linear model corresponding to an arbitrary estimable contrast.

A.1 Vector spaces

Consider a general $n \times p$ matrix X, made up of the p columns x_i. Pre-multiplication by X takes a vector from $\mathbb{R}^p$ to $\mathbb{R}^n$: $w = Xv$. In general only a subspace of $\mathbb{R}^n$ may be reached, consisting of all the vectors which can be constructed as linear combinations of the columns of X, and hence the dimension of this subspace is the number of linearly independent columns of X. This subspace is known as the range or column-space of X, and denoted $\mathbf{C}(X)$, and its dimension is called the rank of X. Clearly $r = \text{rank}(X) \leq p$, and in the case of equality we say X has full column rank. The number of linearly independent rows $\text{rank}(X^T)$ can be shown to equal the number of linearly independent columns [1], and hence $r \leq n$ is also true.

If $r < n$, then $n - r$ of the rows are linearly dependent, i.e. they can be made up from linear combinations of the other rows. This means that in some linear combination of all the rows $w^T X$, we can find w such that the dependency allows us to cancel some rows with others, giving $w^T X = 0$, and the dimension of the subspace of all such w is given by $n - r$. This subspace is known as the left null space of X; every vector w within it is orthogonal to every vector u in $\mathbf{C}(X)$ because $w^T u = w^T(Xv) = (w^T X)v = 0$, and we therefore say that this subspace is the orthogonal complement of the column space, which we denote $\mathbf{C}(X)^\perp$. Because the bases of these two spaces are orthogonal, together, they include $\text{rank}(X) + n - \text{rank}(X) = n$ linearly independent vectors, and therefore span the whole of $\mathbb{R}^n$.

There are naturally an equivalent pair of orthogonal subspaces within $\mathbb{R}^p$ given by the column space of X^T, or the row space of X, and the left null space of X^T, which we call simply the null space of X, $\mathbf{N}(X)$. Any vector $v \in \mathbb{R}^p$ can be written as the sum of a vector in the null space and another vector in the row space, and the action of pre-multiplication with X maps the null space component to zero, and the row space component to a vector in the column space. There is actually a one-to-one mapping from the row space to the column space, which are both rank(X)-dimensional, and the pseudo-inverse which will be introduced below inverts this mapping [1]. These four subspaces are the fundamental subspaces of the matrix X; to summarise their properties:

$$\begin{aligned}
\mathbf{C}(X)^\perp &= \mathbf{N}(X^T) \\
\mathbf{C}(X) &= \mathbf{N}(X^T)^\perp \\
\mathbf{C}(X) \cup \mathbf{N}(X^T) &= \mathbb{R}^n \\
\mathbf{C}(X^T)^\perp &= \mathbf{N}(X) \\
\mathbf{C}(X^T) &= \mathbf{N}(X)^\perp \\
\mathbf{C}(X^T) \cup \mathbf{N}(X) &= \mathbb{R}^p.
\end{aligned}$$

Considering the rank(X) linearly independent columns of X, we may choose these vectors so that they are mutually orthogonal unit vectors, i.e. an orthonormal basis for $\mathbf{C}(X)$, and we may collect them into a matrix U, such that $U^TU = I$. Any vector in $\mathbf{C}(X)$ can be represented as $u = Uv$, and we observe that the matrix $P = UU^T$ maps these vectors onto themselves $Pu = UU^TUv = U(U^TU)v = Uv = u$. Furthermore, due to the orthogonality of the subspaces, any vector in the left null space will be mapped to zero, and hence any general vector $y \in \mathbb{R}^n$, which has components in the column space and left null space, will be mapped solely to its component in the column space, i.e. P projects onto the column space. Because the (left null space) component removed in the projection is orthogonal to the (column) space projected to, P is called the perpendicular projection matrix onto $\mathbf{C}(X)$ [2]. P is symmetric ($P^T = (UU^T)^T = UU^T = P$) and idempotent ($PP = UU^TUU^T = UU^T = P$) and, interestingly, given the non-uniqueness of the basis U, is unique.[1]

The projection matrix P has rank(X) eigenvalues of one, corresponding to vectors in the column space (e.g. the orthonormal vectors in U), and $n-\text{rank}(X)$ eigenvalues of zero, corresponding to vectors in the left null space. This means $\text{rank}(X) = \text{rank}(P) = \text{tr}\,(P)$, since the trace of any matrix is equal to the sum of its eigenvalues.

Any vector $y \in \mathbb{R}^n$ can be written as $Iy = (P+I-P)y = Py+(I-P)y$, since Py is the component in the column space, $(I-P)y$ must be the orthogonal component in the left null space (and indeed, these components are orthogonal since $(Py)^T(I-P)y = y^TP^T(I-P)y$

[1] To prove the uniqueness, hypothesise the existence of a different projection matrix Q; for any vector $y \in \mathbb{R}^n$ we have $Py = Qy$, since they project to the same vector (the component of y in the column space); however, there is no restriction on y, so we may for example choose a vector with zeros everywhere apart from a single one in the i^{th} element, such a vector simply picks out the i^{th} column of Py and of Qy, implying that these columns are equal, and hence, given the free choice of i, that all columns of P and Q are the same.

and $P^T(I-P) = P(I-P) = P - P = 0$). The matrix $R = I - P$ is symmetric and idempotent, and is the unique perpendicular projection matrix onto $\mathbf{C}(X)^{\perp}$.

An orthonormal basis U_n for $R = U_n U_n^T$, can be found from the $n - \text{rank}(X)$ linearly independent vectors in the left null space. The orthogonality of the subspaces means that U and U_n would be mutually orthogonal, and hence that $U_f = [U\ U_n]$ satisfies $U_f^T U_f = I$, and also U_f spans the whole of $\mathbb{R}^n$, meaning that its perpendicular projection matrix $U_f U_f^T$ is also an identity.

We can similarly find orthonormal matrices V, V_n and $V_f = [V\ V_n]$, such that $V_f^T V_f = I$ and $V_f V_f^T = I$, where V and V_n are respectively bases for the row space and null space of X.

A.2 The Singular Value Decomposition

It can be shown [1] that these bases can be chosen such that $X = U_f S_f V_f^T = USV^T$, where S is a (unique)[2] diagonal matrix with $\text{rank}(X)$ positive values on the diagonal, known as the singular values, and S_f is an $n \times m$ matrix with S as its upper-left block and zeros elsewhere. $X = U_f S_f V_f^T$ is known as the singular value decomposition (SVD) of X, and $X = USV^T$ is the compact SVD.[3]

A.2.1 Numerical precision

Among the strengths of the SVD is that has good numerical properties [1]. In particular, the orthogonal matrices U and V preserve lengths of vectors they multiply, while the matrix of singular values S clearly characterises the relative scalings involved for the original matrix. The SVD provides a good measure of the 'effective rank' of a matrix in the case that some rows or columns are almost linearly dependent. Such 'ill-conditioned' matrices might appear to have different rank depending on the exact operation, algorithm or computer platform. By considering singular values below a certain threshold to be equivalent to zero, the resulting modified decomposition approximates in the original matrix in an optimal way (in terms of the Frobenius norm of the difference) and has a clear and numerically stable rank.[4] Furthermore, the condition number of a matrix (the ratio of its largest to smallest singular values) can be used to bound the error of a solution that uses its inverse [1].

A.2.2 Related eigen-decompositions

We find that $XX^T = U_f S_f S_f^T U_f^T$ and $X^T X = V_f S_f^T S_f V_f^T$, which are both in the form of eigen-decompositions for symmetric matrices. $S_f S_f^T$ is an $n \times n$ matrix, while $S_f^T S_f$ is

[2] The matrices U and V are not unique, for example columns corresponding to the same singular value can be negated without changing the product or the spaces spanned. More general results regarding equivalent SVDs can be found in [3].

[3] We use the subscript f on the full SVD instead of the more obvious approach of using a subscript r on the reduced SVD, because the derivations later make extensive use of the reduced form, and hence look less cluttered without the subscript.

[4] For example, MATLAB defines the rank as the number of singular values above a tolerance t, given by $t = \max(m,n) \times \max_i s_i \times \epsilon$ where ϵ is the machine precision ($2.2e^{-16}$ for 64-bit IEEE floating point).

$m \times m$. However, they both have S^2 as their upper-left blocks, implying that the non-zero eigenvalues of X^TX and XX^T are the same. We can also observe

$$XV = USV^TV = US$$
$$U^TX = U^TUSV^T = SV^T$$

which allow us to derive either set of singular vectors corresponding to non-zero singular values from the other set together with S.

A.2.3 The SVD of a projection matrix

A general projection matrix P is symmetric and idempotent, so the eigen-decompositions discussed above, PP^T and P^TP, are both decompositions of P, meaning that U and V in the compact SVD of P are the same.[5] Futhermore, a projection matrix has eigenvalues of either zero or one, meaning the compact SVD's $S = I$. This gives the compact SVD as $P = UU^T$, where U is an $n \times \text{rank}(P)$ matrix. Note that we previously defined a particular projection matrix, onto the space spanned by an orthonormal basis U, as UU^T; here we have shown that the SVD allows us to recover an orthonormal basis for a general projection matrix.

A.3 The Moore-Penrose Pseudoinverse

This section presents some material relating to generalised inverses, and more specifically the Moore-Penrose pseudoinverse, which is useful elsewhere in the thesis.

For a full rank square matrix X, there is a unique inverse X^{-1} such that $X^{-1}X = I = XX^{-1}$. For a 'thin' rectangular matrix of full column rank, i.e. an $n \times m$ matrix X with $\text{rank}(X) = m < n$, X^TX is an $m \times m$ matrix with rank m [1], and is hence invertible, which means we can find a 'left-inverse' $X^L = (X^TX)^{-1}X^T$ such that $X^LX = I$. Similarly, a 'wide' rectangular matrix with full row rank has a right-inverse, $X^R = X^T(XX^T)^{-1}$ such that $XX^R = I$.

For any rank-deficient matrix X (whether square, thin or wide), we may find a non-unique generalised inverse X^- such that $XX^-X = X$ [4]. The Moore-Penrose pseudo-inverse X^+ is the unique generalised inverse which satisfies three additional properties:

$$X = XX^+X$$
$$X^+ = X^+XX^+$$
$$(XX^+)^T = XX^+$$
$$(X^+X)^T = X^+X.$$

The pseudo-inverse can be computed in practice via the compact singular value decomposition $X = USV^T$.[6] We show that $X^+ = VS^{-1}U^T$ satisfies all four necessary

[5] Since $\mathbf{N}(P)$ and $\mathbf{N}(P^T)$ are the same, the full SVD's U_f and V_f can also be chosen to be equal.

[6] Considering small singular values to be zero (see A.2.1), and hence not part of the compact SVD,

properties:

$$
\begin{aligned}
XX^+ &= USV^TVS^{-1}U^T = UU^T \\
XX^+X &= UU^TUSV^T = USV^T = X && \text{(i)} \\
X^+XX^+ &= VS^{-1}U^TUU^T = VS^{-1}U^T = X^+ && \text{(ii)} \\
(XX^+)^T &= (UU^T)^T = UU^T = XX^+ && \text{(iii)} \\
X^+X &= VS^{-1}U^TUSV^T = VV^T \\
(X^+X)^T &= (VV^T)^T = VV^T = X^+X. && \text{(iv)}
\end{aligned}
$$

We can also prove some useful additional identities. Since $X^T = VSU^T$ is already in the form of a compact SVD, its pseudoinverse is $US^{-1}V^T$ which is $(VS^{-1}U^T)^T = (X^+)^T$, showing that the pseudo-inverse and transpose operators commute. We can similarly show that $(X^+)^+ = X$, and that

$$
\begin{aligned}
(X^TX)^+ &= (VS^2V^T)^+ \\
&= VS^{-2}V^T \\
&= X^+(X^T)^+ = X^+(X^+)^T && \text{(A.1)} \\
(X^TX)^+X^T &= VS^{-2}V^TVSU^T \\
&= VS^{-1}U^T = X^+ && \text{(A.2)}
\end{aligned}
$$

where (A.2) automatically recovers the left-inverse in the case that it exists, and

$$
\begin{aligned}
(XX^T)^+ &= (US^2U^T)^+ \\
&= US^{-2}U^T \\
&= (X^T)^+X^+ = (X^+)^TX^+ && \text{(A.3)} \\
X^T(XX^T)^+ &= VSU^TUS^{-2}U^T \\
&= VS^{-1}U^T = X^+ && \text{(A.4)}
\end{aligned}
$$

with (A.4) recovering the right-inverse where it exists.

Note that X^+ also reduces to the standard inverse in the case that it exists, since for square full rank X the 'compact' SVD is actually the full SVD, for which $UU^T = I$ and $VV^T = I$ and therefore $USV^TVS^{-1}U^T = I$ and $VS^{-1}U^TUSV^T = I$ implying $VS^{-1}U^T = X^+ = X^{-1}$. Also, it is interesting to note that a projection matrix is its own pseudo-inverse, which can be verified by checking the basic properties, or observed from the fact that its compact SVD has an identity matrix for S.

Note that (A.1) and (A.3) do not imply that $(AB)^+ = B^+A^+$, which holds only in some special cases [5]; a necessary and sufficient condition is

$$
(A^+A)BB^TA^TA(BB^+) = BB^TA^TA. \tag{A.5}
$$

provides a numerically stable pseudo-inverse.

We observe that $XX^+ = UU^T = P$ — the perpendicular projection matrix onto $\mathbf{C}(X)$, but note that the SVD furnishes a means of computing $P = UU^T$, that is more computationally efficient than computing the pseudo-inverse. We will later also require the projection matrix for $\mathbf{C}(X)^\perp$, which is $R = I - P$. Again, the full SVD allows us to directly compute $R = U_n U_n^T$. While this is no more efficient in itself, it is useful for implementing an efficient permutation test, as described in section D.2.1.

A.4 The General Linear Model

In this section we derive the common t- and F-statistics, and their less commonly considered multivariate generalisation, for contrasts of parameters in a general linear model.

We begin with maximum likelihood solutions for the parameters, and then show that the likelihood ratio principle leads to test statistics of a certain form. Distributional results then motivate the definition of the F-statistic as the most convenient rearrangement of this form. The two-tailed and single-tailed t-statistics are seen to be a simple special case of the F-statistic. Some helpful alternative formulations and extensions are then considered. Results are related to the multivariate case where appropriate.

A.4.1 Notation

There is no universally adopted standard for statistical notation, though some aspects are reasonably consistent. In keeping with [1] and [2], no distinction is made here between scalars and vectors, though we will denote matrices using capital letters. For this reason, we use B and b for the parameters in a multivariate or univariate model,[7] instead of β, which is more common within the literature for univariate data. Following [2], random variables will not be highlighted, since these should be clear from the context.

It is essential to distinguish between the additive Gaussian *errors* in the linear model, denoted as ε ($\mathcal{E}$ in the multivariate case), and the least-squares *residuals* from the fitted model, denoted e or E. For example, the (multivariate) model assumed is $Y = XB + \mathcal{E}$, the model fitted is $\hat{Y} = X\hat{B}$ and the 'error' in this fitted model is the matrix of residuals $E = Y - X\hat{B}$.

Notational clashes have been avoided where confusion may occur. For example, P is reserved for projection matrices, while permutation matrices are written S, for 'shuffle'. However, S has also been used in the singular value decomposition, as there should be little risk of confusion between permutation and the SVD.

A.4.2 Maximum Likelihood for the Multivariate GLM

Consider n independent observations of an m-variate random column vector. The complete data, given by the n-by-m matrix $Y = [y_1\ y_2\ \dots\ y_n]^T$, is modelled as a linear combination

[7] To avoid confusion later, we mention here that the dimensionality of a contrast of the parameters may vary independently of the dimensionality of the parameters themselves, i.e. $c^T b$ is scalar, $C^T b$ a column vector, $c^T B$ a row vector, and $C^T B$ a matrix.

of some unknown parameters with additive error,

$$Y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{bmatrix} = XB + \mathcal{E} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} B + \begin{bmatrix} \varepsilon_1^T \\ \varepsilon_2^T \\ \vdots \\ \varepsilon_n^T \end{bmatrix}.$$

X is an n-by-p design matrix and B is the corresponding p-by-m matrix of unknown parameters. We assume $\varepsilon_i \sim \mathrm{N}(0, V)$ with m-by-m positive definite covariance matrix V. The independence of the observations gives $\mathrm{Cov}[\varepsilon_i, \varepsilon_j] = 0$ for $i \neq j$, and means that the complete likelihood factors into a product over the observations (rows of Y):

$$\begin{aligned} p(Y|B,V) &= \prod_{i=1}^{n} \mathrm{N}(y_i^T|x_i^T B, V) = \prod_{i=1}^{n} \mathrm{N}(y_i|B^T x_i, V) \\ &= \prod_{i=1}^{n} (2\pi)^{-m/2} |V|^{-1/2} \exp -\tfrac{1}{2}(y_i - B^T x_i)^T V^{-1} (y_i - B^T x_i) \end{aligned}$$

Denoting the cost function $L(B,V) = -2 \log p(Y|B,V)$, we seek to minimise:

$$\begin{aligned} L &= \sum_{i=1}^{n} \left(m \log(2\pi) + \log|V| + (y_i - B^T x_i)^T V^{-1} (y_i - B^T x_i) \right) \\ &= nm \log(2\pi) + n \log|V| + \sum_{i=1}^{n} (y_i - B^T x_i)^T V^{-1} (y_i - B^T x_i) \\ &= nm \log(2\pi) + n \log|V| + \sum_{i=1}^{n} \mathrm{tr}\left((y_i - B^T x_i)^T V^{-1} (y_i - B^T x_i) \right) \\ &= nm \log(2\pi) + n \log|V| + \sum_{i=1}^{n} \mathrm{tr}\left(V^{-1} (y_i - B^T x_i)(y_i - B^T x_i)^T \right) \\ &= nm \log(2\pi) + n \log|V| + \mathrm{tr}\left(V^{-1} \sum_{i=1}^{n} (y_i - B^T x_i)(y_i - B^T x_i)^T \right) \\ &= nm \log(2\pi) + n \log|V| + \mathrm{tr}\left(V^{-1} (Y - XB)^T (Y - XB) \right). \end{aligned}$$

Now, we wish to expand the quadratic form

$$\begin{aligned} D &= (Y - XB)^T (Y - XB) \\ &= Y^T Y + B^T X^T X B - Y^T X B - B^T X^T Y, \end{aligned}$$

and complete the square in B (with analogy to the scalar case in x, $ax^2 + 2bx = a(x - b/a)^2 - (b/a)^2$). Recalling section A.3, it is trivial to see that

$$(X^+)^T X^T = (X X^+)^T = P \tag{A.6}$$

and that the following therefore hold:

$$\begin{aligned} X = PX &= (X^+)^T X^T X, \\ X^T &= X^T X X^+. \end{aligned} \tag{A.7}$$

We use (A.7) to expand each term so as to contain $X^T X$, and we then complete the square by collecting these terms together:

$$\begin{aligned} & B^T X^T X B - Y^T X B - B^T X^T Y \\ =& B^T X^T X B - Y^T (X^+)^T X^T X B - B^T X^T X X^+ Y, \\ =& (B - X^+ Y)^T X^T X (B - X^+ Y) - Y^T (X^+)^T X^T X X^+ Y. \end{aligned}$$

From equation A.6 we simplify

$$Y^T (X^+)^T X^T X X^+ Y = Y^T X X^+ Y$$

giving

$$\begin{aligned} D &= Y^T Y - Y^T X X^+ Y + (B - X^+ Y)^T X^T X (B - X^+ Y) \\ &= Y^T R Y + (B - \hat{B})^T X^T X (B - \hat{B}) \end{aligned}$$

where we have defined $\hat{B} = X^+ Y$ and $R = I - P = I - X X^+$, and we note that R is a perpendicular projection matrix, which projects onto the space orthogonal to $\mathbf{C}(X)$, forming the mulitivariate residuals $E = Y - X\hat{B} = RY$.

The cost function may now be written

$$\begin{aligned} L &= nm \log(2\pi) + n \log |V| + \operatorname{tr}\left(V^{-1} Y^T R Y + V^{-1} (B - \hat{B})^T X^T X (B - \hat{B}) \right) \\ L &= nm \log(2\pi) + n \log |V| + \operatorname{tr}\left(V^{-1} Y^T R Y \right) + \operatorname{tr}\left(X (B - \hat{B}) V^{-1} (B - \hat{B})^T X^T \right) \end{aligned}$$

the last term is the trace of a quadratic form with positive definite V^{-1}, and therefore reaches its minimum at $B = \hat{B}$, independently of V. This gives the maximum likelihood solution for the parameters,[8] and leaves us to minimise the remaining terms with respect to V, or equivalently with respect to V^{-1}. Using some standard matrix calculus results[9]

$$\begin{aligned} \frac{\partial \log |V|}{\partial V} &= (V^T)^{-1}, \\ \frac{\partial \operatorname{tr}(VA)}{\partial V} &= A^T, \end{aligned}$$

[8] $\hat{B}$ is also the least squares estimator of the parameters, and the best linear unbiased estimator [2].

[9] These were taken from `http://www.cs.toronto.edu/~roweis/notes/matrixid.pdf`; related results with a slightly more formal notation can be found in [6].

we have

$$L(\hat{B}, V) = l(V^{-1}) = nm\log(2\pi) - n\log|V^{-1}| + \operatorname{tr}\left(V^{-1}Y^T RY\right)$$
$$\frac{\partial l}{\partial V^{-1}} = -nV + Y^T RY,$$

and, setting the gradient to zero, we find the maximum likelihood estimate of the covariance matrix, $\hat{V} = Y^T RY/n$, and hence the value of the likelihood at its maximum:

$$L(\hat{B}, \hat{V}) = nm\log(2\pi) + n\log|Y^T RY| + \operatorname{tr}(nI)$$
$$p(Y|\hat{B}, \hat{V}) = \exp -L/2 = (2\pi)^{-nm/2}|Y^T RY|^{-n/2}e^{-nm/2}. \tag{A.8}$$

A.4.3 The Likelihood Ratio Test

The Neymann-Pearson Lemma states that in order to compare two simple hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, a test based on the ratio of the values of the likelihood under these hypotheses, which rejects H_0 if

$$\frac{p(Y|\theta_0)}{p(Y|\theta_1)} < c$$

for some constant $c \geq 0$, will be the most powerful test for a given size α [7].[10] For more general null and alternative hypotheses, $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta$, this optimality property holds approximately for large samples when using the likelihood ratio test criterion [8]:

$$\frac{\max_{\theta\in\Theta_0} p(Y|\theta)}{\max_{\theta\in\Theta} p(Y|\theta)} < c.$$

In the case of the (multivariate) linear model $Y = XB + E$ we wish to test the null hypothesis that a simplified model $Y = X_0 B_r + E$ holds,[11] where we assume $\mathbf{C}(X_0) \subset \mathbf{C}(X)$, or that the hypothesis is 'nested'. To compare different non-nested designs X and Z it is necessary to resort to more general model comparison techniques such as the various information criteria (Akaike's IC, Bayesian IC, etc.) or fully Bayesian model comparison [9].

The reduced model can be equivalently viewed as a restricted design matrix or as a restriction on the parameters B, as will become clearer below. The maximum of the likelihood from (A.8) becomes

$$p(Y|\hat{B}_r, \hat{V}_0) = (2\pi)^{-nm/2}|Y^T R_0 Y|^{-n/2}e^{-nm/2},$$

where $R_0 = I - X_0 X_0^+$ is the residual forming matrix for the reduced model. This gives

[10] The sampling distribution of the likelihood ratio statistic must be known in order to determine c such that the significance level is α.

[11] The notation B_r is used for the parameters in the reduced model, to avoid confusion with B_0 which is used elsewhere to denote parameters corresponding to X_0 in a model with X partioned into X_1 and X_0.

the likelihood ratio statistic (and its scaled version, known as Wilks' Λ [8]) as:

$$\Lambda^* = \left(\frac{|Y^T R Y|}{|Y^T R_0 Y|}\right)^{n/2} \tag{A.9}$$

$$\Lambda = \frac{|Y^T R Y|}{|Y^T R_0 Y|} \tag{A.10}$$

A.4.4 Distributional Results

Because $\mathbf{C}(X_0) \subset \mathbf{C}(X)$, $P_0 = X_0 X_0^+ \in \mathbf{C}(X)$ and hence

$$PP_0 = P_0 = P_0 P \tag{A.11}$$

$$PR_0 = P(I - P_0) = P - P_0 = R_0 - R \tag{A.12}$$

$$RR_0 = (I - P)R_0 = R. \tag{A.13}$$

R and $R_0 - R$ are orthogonal because $R(R_0 - R) = R - R = 0$, and this orthogonality endows the matrices $Y^T R Y$ and $Y^T(R_0 - R)Y$ with independent Wishart distributions (central under the null hypothesis) [10].

In the univariate case, the scalar quadratic forms $y^T R y$ and $y^T(R_0 - R)y$ have independent χ^2 distributions (central under H_0) with degrees of freedom respectively $\mathrm{tr}\,(R) = \mathrm{tr}\,(I - P) = n - \mathrm{rank}(X)$ and $\mathrm{tr}\,(R_0 - R) = \mathrm{tr}\,(P - P_0) = \mathrm{rank}(X) - \mathrm{rank}(X_0)$. Hence by definition of the F-distribution as the distribution of the ratio of two such independent χ^2 variables divided by their degrees of freedom, we see that

$$\begin{aligned}
\Lambda^{-1} - 1 &= \frac{y^T R_0 y}{y^T R y} - 1 \\
&= \frac{y^T(R_0 - R)y}{y^T R y} \\
\frac{\nu_2}{\nu_1}\frac{1-\Lambda}{\Lambda} &= \frac{y^T(R_0 - R)y/\nu_1}{y^T R y/\nu_2}
\end{aligned}$$

is F distributed with $\nu_1 = \mathrm{rank}(X) - \mathrm{rank}(X_0)$ and $\nu_2 = n - \mathrm{rank}(X)$ degrees of freedom [2].

In summary, the test statistic can be written as

$$F = \frac{MS_H}{MS_E} = \frac{SS_H/DF_H}{SS_E/DF_E} = \frac{y^T(R_0 - R)y}{y^T R y} \Big/ \frac{\mathrm{rank}(X) - \mathrm{rank}(X_0)}{n - \mathrm{rank}(X)} \tag{A.14}$$

where the Sums of Squares, and Mean-Squares are labelled for 'hypothesis' or 'error', and we note the restricted or total sums of squares and degrees of freedom are $SS_R = SS_H + SS_E = y^T R_0 y$ and $DF_R = DF_H + DF_E = n - \mathrm{rank}(X_0)$.[12]

[12] The common terminology 'total' makes more sense in the context of ANOVA main effects tests, where the reduced model contains only the mean, and hence $MS_R = \mathrm{V}\,[y]$.

Rao's F approximation

In the multivariate case, the distribution of Λ is much more complicated. In this thesis, the permutation framework is used for multivariate models, allowing direct testing of Λ without knowing (assuming) its parametric distribtion. However, for completeness, we briefly describe a parametric approach. Λ can be transformed to a statistic with an exact F distribution in special cases where: $m = 1$ or $m = 2$ with arbitrary ν_1; or: $\nu_1 = 1$ or $\nu_1 = 2$ with arbitrary m. Rao's F approximation reproduces these exact cases and provides a reasonable approximation in other situations [4].[13]

Rao's F-approximation defines

$$
\begin{aligned}
t &= n - \operatorname{rank}(X_0) \\
k &= t - \frac{\nu_1 + m + 1}{2} \\
\lambda &= \frac{\nu_1 m - 2}{4} \\
s &= \begin{cases} 1 & \text{if } \min(\nu_1, m) = 1 \\ \sqrt{\frac{\nu_1^2 m^2 - 4}{\nu_1^2 + m^2 - 5}} & \text{otherwise} \end{cases} \\
\nu_1^* &= \nu_1 m \\
\nu_2^* &= ks - 2\lambda \\
F &= \frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} / \frac{\nu_1^*}{\nu_2^*}
\end{aligned}
$$

resulting in F approximately (or exactly, as noted above) distributed as $\mathrm{F}(\nu_1^*, \nu_2^*)$.

A.4.5 Estimable contrasts

Our derivation of $\hat{B} = X^+Y = (X^TX)^+X^TY$ above used a property of the (unique Moore-Penrose) pseudo-inverse. In fact, the necessary properties apply to a group of non-unique generalised inverses [4], meaning that $\hat{B} = X^-Y$ is in turn not unique.[14] If the design matrix is full column rank then the square matrix X^TX is full rank and hence invertible, leading to a unique solution $\hat{B} = (X^TX)^{-1}X^TY$. For rank-deficient X, we find that although different choices of generalised inverse lead to different $\hat{B}$, the fitted data, $\hat{Y} = PY = X\hat{B}$, is invariant to these choices. If we replace the projection matrix $P = XX^+ = X(X^TX)^+X^T$ with $P = X(X^TX)^-X^T$, we can use the result that for any generalised inverse G of X^TX, $XGX^T = XG^TX^T$ is invariant to G [11], to show that PY

[13] Symbols have been translated to avoid conflicts elsewhere in this chapter; Rao $\leftrightarrow$ here: $q \leftrightarrow m$, $m \leftrightarrow k$, $p \leftrightarrow \nu_1$.

[14] Strictly, we should not refer to this solution as an estimate or estimator because of its non-uniqueness; some texts also avoid the 'hat' notation for this reason.

is unique. Alternatively, observe

$$\begin{aligned}\mathrm{E}[\hat{Y}] &= \mathrm{E}[PY] = \mathrm{E}[X\hat{B}] \\ &= \mathrm{E}[XX^{-}Y] = XX^{-}\mathrm{E}[Y] \\ &= XX^{-}XB = XB,\end{aligned}$$

independently of the choice of generalised inverse.

Since we can uniquely estimate XB, we can naturally estimate expressions of the form K^TXB. Therefore a linear compound of the unknown parameters C^TB for a column vector or matrix C is 'estimable' if and only if we can write $C^T = K^TX$ — i.e. a linear combination of the rows of the design matrix. It is common in the neuroimaging community to refer to such an estimable linear function as a 'contrast' [12], and we adopt this terminology here, although in the more general statistics literature a contrast is usually defined as a linear compound whose weights sum to zero [13].

The focus here has been on the conventional definition of estimability, which is concerned only with the theoretical presence of the contrast in the row-space of the design. In practice, machine precision must be taken into account, as discussed briefly in A.2.1, for example considering a contrast to be estimable if its projection perpendicular to the row-space is sufficiently small with respect to numerical precision. More practically still, Smith et al. [14] argue that estimability and efficiency are more meaningful when taking into account the noise or variability, i.e. considering whether a contrast can be estimated with sufficient power to be experimentally useful.

A null hypothesis specified by a contrast $C^TB = 0$ places a restriction on the space in which the solution lies, and hence can be equivalently considered via a restricted design matrix, as employed above to derive the test statistic. It is only the space spanned by the restricted design that matters, and not the exact reduced model X_0, which is non-unique (even for full column rank X). In the following subsections an argument based on subspaces is used to derive an equivalent form of the test statistic (A.14) in terms of the contrast.

The hypothesis subspace

Considering the univariate case,[15] $y^TRy = (Ry)^T(Ry) = e^Te = |e|^2$, and hence the test statistic is function of the squared lengths of the residual vectors e and $e_0 = R_0Y$. The residual vectors lie in vector subspaces perpendicular to the space spanned by their design

[15] In the multivariate case, the determinants of the form $|Y^TRY|$ are not so easily interpreted, but the interpretation of the spaces of $\mathbf{C}(X)$ etc. remains the same.

matrices. Considering the numerator of the test statistic,

$$\begin{aligned} SS_H &= y^T(R_0 - R)y \\ &= y^T P_h y \\ &= |h|^2 \\ P_h &= R_0 - R = P - P_0 \\ &= PR_0 = R_0P = PR_0P = R_0PR_0 \qquad \text{(A.15)} \\ h &= PR_0y \\ &= R_0Py \\ &= R_0X\hat{B}. \qquad \text{(A.16)} \end{aligned}$$

Where (A.15) continues from the manipulations in (A.12). The hypothesis sum of squares is therefore the squared length of a vector which lies in a subspace both orthogonal to that of X_0 and within that of X, i.e. $h \in \mathbf{C}(X_0) \cap \mathbf{C}(X)$ which we denote the orthogonal complement of $\mathbf{C}(X_0)$ with respect to $\mathbf{C}(X)$: $h \in \mathbf{C}(X_0)^\perp_{\mathbf{C}(X)}$ [2]. For finite dimensional spaces, the orthogonal complement of the orthogonal complement returns the original space [2], i.e.

$$\Gamma = \mathbf{C}(X_0)^\perp_{\mathbf{C}(X)} \iff \Gamma^\perp_{\mathbf{C}(X)} = \mathbf{C}(X_0). \qquad \text{(A.17)}$$

In terms of the perpendicular projectors, $P - P_0$ projects onto $\mathbf{C}(X_0)^\perp_{\mathbf{C}(X)}$, and we observe $P - (P - P_0) = P_0$ projects onto the original space $\mathbf{C}(X_0)$.

As an aside, note that (A.16) gives $SS_H = \hat{B}^T X^T R_0 X \hat{B}$ (valid also in the multivariate case), which provides a computationally efficient way of calculating SS_H from previously estimated $\hat{B}$ images and a suitable $p \times p$ matrix $X^T R_0 X$ [15]. The following section effectively provides a formula for directly computing this matrix from a hypothesis $C^T B = 0$.

The test statistic for a contrast

We seek a null hypothesis equivalent to $\mathrm{E}[Y] = X_0 B_r$ or $\mathrm{E}[Y] \in \mathbf{C}(X_0)$ of the form: $C^T B = K^T \mathrm{E}[Y] = 0$ in conjunction with $\mathrm{E}[Y] = XB$. It is apparent that $\mathrm{E}[Y] \in \mathbf{C}(K)^\perp \cap \mathbf{C}(X)$, but this is unhelpful because we have not ensured that $\mathbf{C}(K)$ is a subset of $\mathbf{C}(X)$. The key is to note that the contrast can equivalently be written:

$$\begin{aligned} C^T B &= 0 \\ &= K^T XB \\ &= K^T PXB \\ &= (PK)^T \mathrm{E}[Y] \end{aligned}$$

where $PK \in \mathbf{C}(X)$ gives

$$\begin{aligned} \mathrm{E}[Y] &\in \mathbf{C}(PK)^{\perp} \cap \mathbf{C}(X) \\ &\in \mathbf{C}(PK)^{\perp}_{\mathbf{C}(X)} \end{aligned}$$

The equivalence of $\mathbf{C}(PK)^{\perp}_{\mathbf{C}(X)}$ with $\mathbf{C}(X_0)$ and (A.17) imply $\mathbf{C}(PK) = \mathbf{C}(X_0)^{\perp}_{\mathbf{C}(X)}$, and therefore that the projection matrix $P_h = R_0 - R$ is equivalent to projection onto the space spanned by PK, implying

$$P_h = P_{PK} = PK(PK)^{+} = PK(K^T PK)^{+} K^T P. \tag{A.18}$$

Recalling $K^T X = C^T$, $P = X(X^T X)^{+} X^T$ and $PY = X\hat{B}$, we rewrite:

$$\begin{aligned} SS_H &= Y^T P_h Y \\ &= Y^T PK(K^T PK)^{+} K^T PY \\ &= \hat{B}^T X^T K(K^T X(X^T X)^{+} X^T K)^{+} K^T X\hat{B} \\ &= \hat{B}^T C(C^T (X^T X)^{+} C)^{+} C^T \hat{B}. \end{aligned} \tag{A.19}$$

Since each independent column of the contrast C places one restriction on the form of X_0, we also have $DF_H = \nu_1 = \mathrm{rank}(X) - \mathrm{rank}(X_0) = \mathrm{rank}(C)$, and therefore the hypothesis mean-square is expressed only in terms of the contrast and the original design matrix. Finally, for the univariate case, we may write equation (A.14) as:

$$\begin{aligned} F &= \frac{MS_H}{MS_E} \\ &= \frac{\hat{b}^T C(C^T (X^T X)^{+} C)^{+} C^T \hat{b}}{Y^T RY} \Big/ \frac{\mathrm{rank}(C)}{n - \mathrm{rank}(X)} \\ &= \frac{\hat{b}^T C(C^T (X^T X)^{+} C)^{+} C^T \hat{b}/\mathrm{rank}(C)}{MS_E} \end{aligned} \tag{A.20}$$

Note that the expression for SS_H, equation A.19, also gives the sums of squares and products matrix in the multivariate case.

In the univariate case, the overall expression for the F-statistic can also be motivated in terms of a Wald pivot for the contrast [16]. Given $y \sim \mathrm{N}(Xb, \sigma^2 I)$, the basic properties of expectation give $C^T\hat{b} = C^T X^{+} Y \sim \mathrm{N}(C^T X^{+} Xb, \sigma^2 C^T X^{+} (X^{+})^T C)$. For an estimable contrast, $C^T X^{+} Xb = K^T X X^{+} Xb = K^T Xb = C^T b$ as expected (i.e. the natural estimator of an estimable compound is unbiased). Using identity (A.1) we have $C^T\hat{b} \sim \mathrm{N}(C^T b, \sigma^2 C^T (X^T X)^{+} C)$. From this multivariate normal distribution, the squared Mahalanobis distance of the estimate from the null hypothesis of $C^T b = 0$ is given by:

$$\hat{b}^T C(C^T (X^T X)^{+} C)^{+} C^T \hat{b}/\sigma^2,$$

which is simply a scaled version of SS_H, or equivalently, a scaled version of the F-statistic if σ^2 is replaced with the estimate MS_E.

The t-statistic

Considering univariate data and a single column vector contrast $C = c$, equation (A.20) simplifies because $c^T\hat{b}$ and $c^T(X^TX)^+c$ are scalar, giving:

$$F = \frac{(c^T\hat{b})^2}{MS_E\, c^T(X^TX)^+c},$$

which is the square of

$$t = \frac{c^T\hat{b}}{\sqrt{\hat{\sigma^2}\, c^T(X^TX)^+c}}. \tag{A.21}$$

The above shows the equivalence between the two-tailed t-test and the F-test for the same contrast. Since $c^T\hat{b}$ is a scalar with the desired sign of a one-tailed contrast, we may implement a one-sided test with the above equation or using the full and reduced models form (A.14) with the sign of $\sqrt{F}$ set to match that of $c^T\hat{b}$.

A.4.6 Extended hypotheses

We will now consider testing some different forms of null hypothesis, with the purpose of drawing attention to some less commonly used tests, which are unavailable in some standard neuroimaging statistics software (such as SPM and FSL).

Following the subsection above, one might observe that the standard t-statistic for a mean is valid for testing departure from any hypothesised value, not necessarily zero, in which case the numerator of (A.21) would be $\hat{b}-\bar{b}$ for scalar $b = \bar{b}$ under the null hypothesis. This generalises to contrasts of a vector b, and of a matrix in the multivariate case [10]. For an estimable contrast $C^T = K^TX$, to test the null hypothesis $C^TB = C^T\bar{B}$ one can use either of the following forms of SS_H in the (possibly multivariate) test statistics:

$$\begin{aligned} SS_H &= (\hat{B} - \bar{B})^TC(C^T(X^TX)^+C)^+C^T(\hat{B} - \bar{B}) \\ &= (Y - X\bar{B})^T(R_0 - R)(Y - X\bar{B}). \end{aligned}$$

The expressions may also be used for hypotheses of the form $C^TB = \gamma$ where $C^T\bar{B} = \gamma$ has at least one solution $\bar{B}$ [10].

In the multivariate case two further extensions may be of interest. For a null hypothesis of the form $C^TBD = \bar{B}$, where D is a $p \times q$ matrix with $\text{rank}(D) = q < p$ one can consider the transformed model $YD = XBD + \mathcal{E}D$ [10], because the covariance matrix of the vectorised $\mathcal{E}D$ retains its block diagonal structure ($\text{Cov}[\varepsilon_iD, \varepsilon_jD] = 0$ for $i \neq j$). Hence the sums of squares and products are simply modified to:

$$\begin{aligned} SS_H &= (\hat{B}D - \bar{B}D)^TC(C^T(X^TX)^+C)^+C^T(\hat{B}D - \bar{B}D) \\ &= (YD - X\bar{B})^T(R_0 - R)(YD - X\bar{B}), \\ SS_E &= (YD)^TRYD. \end{aligned}$$

Secondly, one might ask whether a subset of the dependent variables contains all the

significantly useful information about XB. Partioning the model as:

$$[Y_A\ Y_B] = X[B_A\ B_B] + [\mathcal{E}_A\ \mathcal{E}_B],$$

the conditional distribution of Y_B given Y_A is itself a multivariate linear model of the form:

$$Y_B = Y_A\Gamma + X\Delta + \mathcal{E}_{B|A}.$$

Considering the above as a standard multivariate linear model with the design partitioned into interest X and nuisance Y_A, a significance test of Δ can be used to determine the importance of the subset Y_B [10] (pp.64–68).

A.4.7 Explicit forms for X_0 and X_h

Equation (A.20) allows for the computation of the F-statistic from a contrast, but we have not yet given an explicit reduced design matrix X_0 (recall that this is not unique). One such option, motivated from $\mathbf{C}(X_0) = \mathbf{C}(PK)^{\perp}_{\mathbf{C}(X)}$, is $X_0 = R_{PK}X$, with $R_{PK} = I - PK(PK)^+$ as usual [2].

Another option, more common in the SPM literature, derives from the following argument: the constraint $C^TB = 0$ restricts the parameters to be in the left null space of the contrast [1], for which the perpendicular projection matrix is $I - CC^+$, and the reduced model projecting the parameters into this restricted space can be seen to be $Y = X(I - CC^+)B + \mathcal{E}$, which immediately gives $X_0 = X(I - CC^+)$. The disadvantage of this formulation, is that it is not at all obvious how to relate the two equivalent forms of SS_H:

$$\begin{aligned}
SS_H &= Y^T(R_0 - R)Y = Y^TP - P_0Y \\
&= Y^T(XX^+ - X_0X_0^+)Y \\
&= Y^T(XX^+ - X(I - CC^+)(X(I - CC^+))^+)Y, \\
SS_H &= \hat{B}^TC(C^T(X^TX)^+C)^+C^T\hat{B} \\
&= Y^T(X^+)^TC(C^T(X^TX)^+C)^+C^TX^+Y.
\end{aligned}$$

One may also find several different matrices or bases X_h that span the same space and have the same (unique) perpendicular projection matrix $R_0 - R = P_h = X_hX_h^+$. We have already seen in equation A.18 that PK is one such solution, Kiebel et al. [17] also give R_0XC in their equation 39. Andrade et al. [15] give PR_0X, which can be simplified as $PR_0X = R_0PX = R_0X$. The equivalence of the spaces spanned by these two matrices can be observed from:

$$\begin{aligned}
\mathbf{C}(X) &= \mathbf{C}([XC\ X_0]) \\
\mathbf{C}(R_0X) &= \mathbf{C}([R_0XC\ 0]) \\
&= \mathbf{C}(R_0XC),
\end{aligned}$$

where the last line follows from the fact that appending the matrix of zeros adds nothing to the column space of R_0XC.

To see the equivalence of the spaces spanned by R_0X and PK we show that the projector for R_0X leads to P_h as follows:

$$\begin{aligned}(R_0X)(R_0X)^+ &= (R_0X)(X^TR_0X)^+X^TR_0 \\ &= PR_0X(X^TR_0X)^+X^TR_0P \\ &= X(X^TX)^+X^TR_0X(X^TR_0X)^+X^TR_0X(X^TX)^+X^T \\ &= X(X^TX)^+(X^TR_0X)(X^TR_0X)^+(X^TR_0X)(X^TX)^+X^T \\ &= X(X^TX)^+(X^TR_0X)(X^TX)^+X^T \\ &= PR_0P = P_h.\end{aligned}$$

We may also express PK more directly in terms of X and C by noting the following:

$$\begin{aligned}C^T &= K^TX \\ C &= X^TK \\ (X^T)^+C &= (X^T)^+X^TK \\ &= PK,\end{aligned}$$

where the the last equality results from interchanging the transpose and pseudoinverse operations and using (A.6). $X_h = (X^T)^+C$ is Poline et al.'s H, though their derivation seems less clear [18].

A.4.8 Partitioned reparameterisation

We showed above that a suitable reduced design matrix is $X_0 = XC_0$ where $C_0 = I - CC^+$. It might reasonably be assumed that the partitioned matrix $X_p = [X_1\ X_0]$ with $X_1 = XC$ would give the same results for a contrast $C_p^T = [I_{r_1}\ 0_{r_1 \times r_0}]$,[16] as the original design X would for the contrast C. It turns out that this is true in terms of the spaces spanned by the full and reduced models, and therefore in terms of the test statistics. Suprisingly however, it is not true for the estimate of the contrast itself: $C_p^TX_p^+Y \neq C^TX^+Y$. This property can be important in some circumstances, for example Smith et al. [14] required it for their characterisation of design efficiency. Beckmann et al. [19] showed in their appendix B that a suitable partitioned model can be found. They assumed full column rank X, and derived rather complicated expressions. We can simplify the expressions slightly by defining a function f such that

$$f(C) = (X^TX)^{-1}C(C^T(X^TX)^{-1}C)^{-1}.$$

[16] r_1 and r_0 are respectively the number of interest and nuisance covariate columns in the partitioned design. Typically, the interest will have full column rank, $r_1 = \text{rank}(X_1) = \text{rank}(C)$, but the nuisance will be rank-deficient, with $r_0 \geq \text{rank}(X_0)$. For example, $X_0 = XC_0$ will have $r_0 = p$.

The expressions from [19] can then be written

$$X_1 = Xf(C)$$
$$X_0 = Xf((I - Cf(C)^T)C_2),$$

where C_2 is a matrix such that $[C\ C_2]$ is square and full rank (and hence invertible). C_2 may be taken as the basis for the left null space (e.g. from the SVD) of C, and hence is related to the previously defined $C_0 = I - CC^+$ by $C_0 = C_2C_2^+$.

Here, we extend the results of [19] to general rank-deficient X, and we find a simpler and more intuitive expression for X_p. We start by defining the matrix $\Gamma^T = [C\ C_0]$. Note that $C_0 = I - CC^+$ is a $p \times p$ symmetric idempotent projection matrix, with $\text{rank}(C_0) = p - \text{rank}(C)$. Furthermore, C_0 and C are orthogonal since $C_0^TC = C_0C = (I - CC^+)C = C - CC^+C = C - C = 0$, which means that the number of linearly independent rows in Γ (and hence its rank) is the sum of the ranks of C and C_0, which is p, showing that Γ has full column rank.

For a partitioned matrix with orthogonal submatrices such as Γ, it can be shown that the Moore-Penrose pseudo-inverse consists of the pseudo-inverses of the blocks [6]: $\Gamma^+ = [(C^T)^+\ (C_0^T)^+]$.[17] In this case, $(C_0^T)^+ = C_0$ thanks to its status as a projection matrix.

Note that ΓX^+Y contains the contrast of interest C^TX^+Y in its first $\text{rank}(C)$ rows. If we can show that Γ and $Z = X^+$ satisfy the condition of (A.5), then we have $\Gamma Z = (Z^+\Gamma^+)^+$ where the matrix

$$\begin{aligned} X_p = Z^+\Gamma^+ &= X\Gamma^+ \\ &= X[(C^T)^+\ C_0] \\ &= [X(C^T)^+\ X_0] \end{aligned} \tag{A.22}$$

therefore produces the contrast of interest in the first $\text{rank}(C)$ rows of X_p^+Y. Furthermore, we can verify that X_p spans the same space as X, as required, because

$$\begin{aligned} X_pX_p^+ &= X[(C^T)^+\ C_0]\Gamma Z \\ &= X[(C^T)^+\ C_0][C\ C_0]^TX^+ \\ &= X((C^T)^+C^T + C_0)X^+ \\ &= X((CC^+)^T + I - CC^+)X^+ \\ &= X(CC^+ + I - CC^+)X^+ \\ &= XX^+. \end{aligned}$$

[17] We have observed that the following seems to hold: $[X_1\ X_0]^+ = [((R_0X_1)^+)^T\ ((R_1X_0)^+)^T]^T$, which demonstrates the simpler form if the blocks are orthogonal ($R_iX_j = X_j - P_iX_j = X_j$), as well as helping to clarify that estimated $\hat{B}$ are based on what is explained uniquely by each block, i.e. for the orthogonal parts of the blocks, and hence the interest $\hat{B}_1$ are unchanged by explicit orthogonalisation of X_1 with respect to X_0. However, this identity is not present in an obvious form in [6], and we have so far been unable to prove it.

Hence, the final step in proving the validity of the expression in (A.22), which has replaced $X_1 = XC$ with $X_1 = X(C^T)^+$, is to show that the condition of (A.5) does indeed hold. We begin by noting that because Γ is full column rank it has a left-inverse, or equivalently $\Gamma^+\Gamma = I$. Equation (A.5) therefore requires that the matrix $ZZ^T\Gamma^T\Gamma$ is invariant to post-multiplication by $ZZ^+ = X^+X$.

$$\begin{aligned}\Gamma^T\Gamma &= [C\ C_0][C\ C_0]^T \\ &= CC^T + C_0^T \\ &= CC^T + I - CC^+ \\ &= CC^T + I - C(C^TC)^+C^T.\end{aligned}$$

At this point, because we are considering an estimable contrast, we have $C^T = K^TX$, and we note that

$$C^TZZ^+ = C^TX^+X = K^TXX^+X = K^TX = C^T, \tag{A.23}$$

demonstrating that all terms of $ZZ^T\Gamma^T\Gamma$ ending with C^T are invariant as required, leaving us only with the term ZZ^T. Using the first property of the pseudo-inverse and noting ZZ^+ is symmetric, we have

$$\begin{aligned}Z &= ZZ^+Z \\ Z^T &= Z^TZZ^+ \\ ZZ^T &= ZZ^TZZ^+\end{aligned}$$

which shows the last remaining necessary invariance.

In comparison to appendix B of Beckmann et al. [19], we must make two further observations. Firstly, Beckmann et al. allowed for non-scalar covariance V using weighted least squares (WLS), whereas we have assumed ordinary least squares (OLS). However, the pre-whitening (using W such that $WW^T = V^{-1}$) that leads to WLS depends only on V, so may be applied to Y and X_p just as easily as to Y and the original X. The second point is that Beckmann's X_1 and X_0 are orthogonal to each other, while this is not true of our expressions. Appendix A of [19] showed that the interest covariates could be orthogonalised with respect to the nuisance without changing the interest parameters or overall residuals. Technically, that proof assumed a full rank X_p, but we have observed (though not proven) that the same appears to be true with either or both X_1 and X_0 rank-deficient.[18]

Slightly more generally, we note (without proof) that adding any amount of any vector in the column-space of the nuisance to either the data and/or the interest covariates changes only the parameter estimates for the nuisance, and pre-multiplication with R_0 can be seen to subtract a linear combination of the columns of X_0 weighted by X_0^+ post-multiplied by the data or interest covariate.

[18] It should be relatively straight-forward to adapt the proof from [19] to use expressions valid for the pseudo-inverse [6], but we have not pursued this here, since the result is already quite intuitive with a subspace interpretation of linear modelling.

A.4.9 Orthogonalised reparameterisation

If we consider the explicitly orthogonalised $X_1^* = R_0 X_1 = R_0 X (C^T)^+$, then $X_p^* = [X_1^* \; X_0]$ has orthogonal partitions, and hence its pseudo-inverse consists of the pseudo-inverses of the partitions, as described above. Hence the interest $\hat{B}_1$ may be computed via $(R_0 X (C^T)^+)^+ Y$ alone, and the corresponding nuisance will become $\hat{B}_0^* = X_0^+ Y$.

Now, consider orthogonalising the data with respect to the nuisance, first in terms of the sums of squares for the original reduced and full models:

$$\begin{aligned} SS_R &= Y^T R_0 Y = (R_0 Y)^T R_0 (R_0 Y), \\ SS_E &= Y^T R Y = Y^T R_0 R R_0 Y = (R_0 Y)^T R (R_0 Y), \end{aligned}$$

where $R = R_0 R R_0$ follows from (A.12) and (A.15). This shows that orthogonalising the data has no effect on the residuals or test statistics.

Next, consider the effect of data-orthogonalisation on the estimated interest parameters using the orthogonalised interest covariates:

$$\begin{aligned} X_1^* Y &= (R_0 X (C^T)^+)^+ Y \\ &= ((R_0 X (C^T)^+)^T R_0 X (C^T)^+)^+ (R_0 X (C^T)^+)^T Y \\ &= (C^+ X^T R_0 X (C^T)^+)^+ C^+ X^T R_0 Y \\ &= X_1^* R_0 Y, \end{aligned}$$

i.e. $\hat{B}_1$ is also unchanged by orthogonalisation of the data. The estimated nuisance parameters for the orthogonalised data are equal to

$$\hat{B}_0^* = X_0^+ R_0 Y = (X_0^T X_0)^+ X_0^T R_0 Y = 0,$$

because $X_0^T R_0 = R_0 X_0 = 0$. This means that the nuisance covariates are no longer needed if both the interest covariates and the data have been orthogonalised using R_0. Hence the regression of Y on $[X_1 \; X_0]$ is equivalent to that of $R_0 Y$ on $R_0 X_1$ alone, in terms of $\hat{B}_1$ and the sums of squares. For parametric equivalence of the t or F statistics, note that the error degrees $DF_E = \mathrm{tr}\,(R_0)$ would be wrong if they were calculated anew for this interest-only model, i.e. the r_0 columns of zeros in $R_0 X_p = R_0 [X_1 \; X_0] = [X_1^* \; 0_{n \times r_0}]$ would need to be counted as $\mathrm{rank}(X_0)$.

A.4.10 Summary of alternative regression models

The following regression models are all equivalent in terms of the estimated interest parameters, contrast, and test statistics (using the notation data:design:contrast)

$$Y : X : C \tag{A.24}$$
$$Y : [X_1 \ X_0] : C_p \tag{A.25}$$
$$Y : [X_1^* \ X_0] : C_p \tag{A.26}$$
$$R_0 Y : X_1^* : I_{r_1}, \tag{A.27}$$

where, as defined earlier (and taking one particular option for X_0)

$$\begin{aligned} X_1 &= X(C^T)^+ \in \mathbb{R}^{n \times r_1}, \\ X_1^* &= R_0 X_1, \\ X_0 &= X(I - CC^+) \in \mathbb{R}^{n \times p}, \\ C_p &= [I_{r_1} \ 0_{r_1 \times p}]^T. \end{aligned}$$

A.4.11 Other reparameterisations

Kiebel et al.'s form of $X_h = R_0 XC$ emphasises that the contrast tests what can be explained by part of the design (XC) over and above what is explained by the null model X_0 (by orthogonalising XC using R_0). Poline et al. [18] therefore argue that to test for all the variance explained by XC, without adjusting for X_0, it is necessary for X_h to be given by XC alone. Using the other form $X_h = PK = (X^T)^+ C$ they show that the contrast $C^F = X^T XC$, achieves the desired result since it gives rise to $X_h^F = (X^T)^+ X^T XC = PXC = XC$. This means that one can use C_F with the estimated parameters from the full model to get results equivalent to those from fitting a new model which has X_0 orthogonalised with respect to $X_1 = XC$.

More generally, Poline et al. [18] state (unfortunately without proof or reference) that the equivalent of a contrast C_a in an alternative transformed model $X_a = XT$ is given by[19]

$$C^T = C_a^T (T^T X^T XT)^+ T^T X^T X.$$

This equation can be simplified from the above form given in [18]:

$$C^T = C_a^T (X_a^T X_a)^+ X_a^T X = C_a^T X_a^+ X. \tag{A.28}$$

Note that this reparameterisation is going in the other direction to that in section A.4.8; this expression does not allow us to determine T or X_a from knowledge of C and C_a.

Importantly, only SS_H is altered by this reparameterisation, and not SS_E, so one cannot directly achieve the equivalent of e.g. simple regression by reparameterising a multiple regression model to contain a single non-constant vector. It should be possible to find a

[19] There is an error in equation 19 of [18], which neglects the transposes on C and C_a.

suitable transformation of multiple F-statistics each resulting from different reparameterisations of the complete model X to achieve this, and in general to test any contrast for any model nested within X without needing to re-fit the new model(s). This might be a useful area for further work.

Bibliography

[1] G. Strang, *Linear Algebra and its Applications*, 3rd ed. Brooks/Cole, Thomson Learning, 1988. ˆ345, 346, 347, 348, 350, 360

[2] R. Christensen, *Plane Answers to Complex Questions: The Theory of Linear Models*, 3rd ed. Springer, 2002. ˆ346, 350, 352, 354, 357, 360

[3] R. Horn and C. Johnson, *Topics in Matrix Analysis*. Cambridge University Press, 1994. ˆ347

[4] C. R. Rao, *Linear Statistical Inference and its Applications*. Wiley, 1965. ˆ348, 355

[5] T. Greville, "Note on the generalized inverse of a matrix product," *SIAM Rev*, vol. 8, no. 4, pp. 518–521, 1966. [Online]. Available: http://www.jstor.org/stable/2027337 ˆ349

[6] J. Schott, *Matrix analysis for statistics*. Wiley, New York, 1997. ˆ352, 362, 363

[7] R. Hogg and A. Craig, *Introduction to mathematical statistics*, 5th ed. Macmillan New York, 1989. ˆ353

[8] R. Johnson and D. Wichern, *Applied multivariate statistical analysis*, 5th ed. Prentice Hall, Upper Saddle River, NJ, 2002. ˆ353, 354

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001. ˆ353

[10] R. Christensen, *Advanced Linear Modeling: Multivariate, Time Series, and Spatial Data; Nonparametric Regression and Response Surface Maximization*, 2nd ed. Springer, 2001. ˆ354, 359, 360

[11] C. E. McCulloch and S. R. Searle, *Generalized, Linear, and Mixed Models*. Wiley, 2001. ˆ355

[12] K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, *Statistical parametric mapping: the analysis of functional brain images*. Academic Press, Elsevier, London, 2007. ˆ356

[13] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, 8th ed. Iowa State University Press, 1989. ˆ356

[14] S. Smith, M. Jenkinson, C. Beckmann, K. Miller, and M. Woolrich, "Meaningful design and contrast estimability in FMRI." *Neuroimage*, vol. 34, no. 1, pp. 127–136, Jan. 2007. ˆ356, 361

[15] A. Andrade, A. L. Paradis, S. Rouquette, and J. B. Poline, "Ambiguous results in functional neuroimaging data analysis due to covariate correlation." *Neuroimage*, vol. 10, no. 4, pp. 483–486, Oct. 1999. ˆ357, 360

[16] M. G. Kenward and J. H. Roger, "Small sample inference for fixed effects from restricted maximum likelihood." *Biometrics*, vol. 53, no. 3, pp. 983–997, Sep. 1997. ˆ358

[17] S. J. Kiebel, D. E. Glaser, and K. J. Friston, "A heuristic for the degrees of freedom of statistics based on multiple variance parameters." *Neuroimage*, vol. 20, no. 1, pp. 591–600, Sep. 2003. ˆ360

[18] J.-B. Poline, F. Kherif, and W. Penny, *Contrasts and Classical Inference*, 2nd ed. Academic Press, 2004, ch. 8. [Online]. Available: http://www.fil.ion.ucl.ac.uk/spm/doc/books/hbf2/pdfs/Ch8.pdf ˆ361, 365

[19] C. F. Beckmann, M. Jenkinson, and S. M. Smith, "General multi-level linear modelling for group analysis in FMRI," Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Tech. Rep. TR01CB1, 2001. [Online]. Available: http://www.fmrib.ox.ac.uk/analysis/techrep/tr01cb1/ ˆ361, 362, 363

Appendix B

The matrix exponential

This appendix presents the matrix generalisations of the exponential and natural logarithm functions, which play a key part in Riemannian analysis of strain tensors and the treatment of matrices in Lie groups, relevant to chapters 4 and 5.

B.1 Matrix powers

The matrix exponential and logarithm can be defined as infinite power series, exactly as for their scalar counterparts. We first note that only for square A may one define $A^2 = AA$, with the obvious extension to higher powers. Diagonalisable A — which includes symmetric positive definite (SPD) matrices — are a useful special case. For $A = UDU^{-1}$ we have, for the square and the general power:

$$A^2 = UDU^{-1}UDU^{-1} = UD^2U^{-1} \tag{B.1}$$

$$A^k = UD^kU^{-1}, \tag{B.2}$$

where the power of the diagonal matrix simply involves the powers of the terms on the diagonal, i.e. in terms of eigenvalues,

$$\lambda(A^k) = \lambda^k(A), \tag{B.3}$$

which can be seen to be true even for non-diagonalisable A, since

$$Av = \lambda v \Rightarrow A^2v = A(\lambda v) = \lambda^2 v.$$

This may further be generalised to fractional and negative powers, i.e. the inverse of a diagonalisable matrix has the same eigenvectors with the reciprocals of the eigenvalues. Note though that not all invertible matrices are diagonalisable; the former property requires only non-zero eigenvalues, while the latter requires a complete set of eigenvectors. To give a simple example, the matrix [1 1; 0 1] has a repeated eigenvalue of 1, and hence is invertible, but it has only one eigenvector ([1; 0]) so it cannot be diagonalised.

B.1.1 Matrix square roots

We observe from (B.2) that a diagonalisable matrix with non-negative eigenvalues has a real matrix square root given by $A^{1/2} = UD^{1/2}U^{-1}$; it can also be shown that a symmetric positive semidefinite (SPS) matrix has a *unique* SPS square root [1]. Square roots for general matrices may be complex and non-unique, and even when real square roots exist they may be difficult to find. Consider a geometric example, a 180° rotation around the z-axis in 3D (denoted $R_z(180)$) can clearly be obtained from the square of a rotation around z by plus or minus 90°. The homogeneous matrix for $R_z(180)$ is `diag([-1 -1 1 1])`, which is already diagonalised (with pairs of repeated eigenvalues, -1 and 1, an identity matrix of eigenvectors). However, this leads to a complex square root, `diag([j j 1 1])`. Instead of the diagonalisation approach, one could attempt to use the Denman-Beavers iteration [2], which, in one form (equation 6.7 of [2]), uses the identity $A^{1/2} = A^{-1/2}A$ to provide an iteration:

$$A_{k+1}^{1/2} = (A_k^{1/2} + A_k^{-1/2}A)/2.$$

This iteration is typically initialised with either $A_0^{1/2} = I$ or $A_0^{1/2} = A$, but unfortunately, either of these leads to the first iteration computing $(A+I)/2$ which is the singular matrix `diag([0 0 1 1])` for $R_z(180)$, preventing further iterations. If the iteration is instead initialised with a 45° rotation around z, then $R_z(90)$ is correctly recovered; initialisation with $R_z(-45)$ converges to $R_z(-90)$.

While we have demonstrated non-uniqueness, and numerical challenges, it is not immediately clear whether any real invertible matrix would fail to have a real square root. Smith et al. [3] state without reference that 'not all affine transformations have exact matrix square roots that are also affine transformations'. We briefly consider a related question in section B.3 below.

B.2 Series definitions

The matrix exponetial function is defined as:

$$\operatorname{expm}(A) = I + A + \frac{A^2}{2} + \cdots = \sum_{k=0}^{\infty} \frac{A^k}{k!}. \tag{B.4}$$

Expression (B.2) allows diagonalisable matrices to be exponentiated simply by exponentiating their eigenvalues. For more general matrices, efficient scaling-and-squaring approaches can be used instead of naïve evaluation of the series [4].

The matrix logarithm may be defined with the following power series [1]:

$$\operatorname{logm}(I - A) = \sum_{k=1}^{\infty} \frac{A^k}{k}, \tag{B.5}$$

which is guaranteed to converge if all of the eigenvalues of A have magnitude strictly less than one. This would imply that the eigenvalues of $B = I - A$ satisfied $|1 - \lambda(B)| < 1$, but the logarithm can be computed for more general matrices, as we shall see next.

The definition of the matrix logarithm as the inverse of the matrix exponential function means that for diagonalisable $(n \times n)$ matrices we need only compute the (scalar) logarithms of their eigenvalues:

$$\operatorname{logm}(A) = \operatorname{logm}\left(UDU^{-1}\right) = U \operatorname{logm}(D) U^{-1} = \sum_{i=1}^{n} u_i \log(\lambda_i) v_i^T, \tag{B.6}$$

where $V = U^{-1}$ and u_i and v_i represent the i^{th} columns of their respective matrices. This implies that all SPD matrices have real and symmetric matrix logarithms (with real, but not necessarily positive eigenvalues), regardless of the magnitude of the matrix's eigenvalues.

In general, the (possibly complex) matrix logarithm of A exists if and only if A is invertible [5]. There can be infinitely many matrix logarithms due to the non-uniqueness of the scalar logarithms of complex eigenvalues.

For real and invertible (but not necessarily diagonalisable) A, if none of the eigenvalues of A are on the negative real line, then A has real logarithms; the principal logarithm is the unique such real logarithm whose complex eigenvalues have imaginary part in the open interval $(-\pi, \pi)$ [6]. Note that this also implies that for an SPD matrix written $A = UDU^T$, there is a unique principal $\operatorname{logm}(A) = U \operatorname{logm}(D) U^T$.

B.3 Properties of the matrix exponential and logarithm

Under a matrix similarity transform of $A \to MAM^{-1}$, both the matrix exponential and logarithm transform in the same way:

$$\operatorname{expm}\left(MAM^{-1}\right) = M \operatorname{expm}(A) M^{-1}, \tag{B.7}$$

$$\operatorname{logm}\left(MAM^{-1}\right) = M \operatorname{logm}(A) M^{-1}. \tag{B.8}$$

This is obvious for the exponential, where the similarity terms cancel in the power series just like the eigenvector terms in (B.1). For the logarithm, the series (B.5) allows a similar argument, noting that

$$B = I - A \to I - MAM^{-1} = M(I - A)M^{-1} = MBM^{-1}.$$

Alternatively, for diagonalisable matrices, it is easy to see that the similarity transform M can be absorbed into the diagonalising similarity U:

$$\operatorname{logm}\left(MAM^{-1}\right) = \operatorname{logm}\left(MUDU^{-1}M^{-1}\right) = \operatorname{logm}\left(CDC^{-1}\right)$$

with $C = MU$, then (B.6) gives

$$\operatorname{logm}\left(CDC^{-1}\right) = C \operatorname{logm}(D) C^{-1} = MU \operatorname{logm}(D) U^{-1}M^{-1} = M \operatorname{logm}(A) M^{-1}.$$

For diagonalisable A, the fact that powers, exponential and logarithm all act on the

eigenvalues gives the following useful relationships:

$$\operatorname{tr}(\operatorname{logm}(A)) = \sum_{i=1}^{n} \log \lambda_i = \log \prod_{i=1}^{n} \lambda_i = \log|A|, \tag{B.9}$$

$$|\operatorname{expm}(A)| = \prod_{i=1}^{n} \exp(\lambda_i) = \exp \sum_{i=1}^{n} \lambda_i = \exp \operatorname{tr}(A); \tag{B.10}$$

$$\begin{aligned} \operatorname{logm}\left(A^k\right) &= U \operatorname{logm}\left(D^k\right) U^{-1} = \sum_{i=1}^{n} u_i \log(\lambda_i^k) v_i^T \\ &= \sum_{i=1}^{n} u_i k \log(\lambda_i) v_i^T = k \operatorname{logm}(A). \end{aligned} \tag{B.11}$$

The last relationship holds for negative and/or fractional powers, meaning that $\operatorname{expm}(A)$ and $\operatorname{expm}(-A)$ are inverses of each other, and that the matrix square root can be easily computed from the logarithm. The existence of the principal logarithm then implies that a real square root exists for any diagonalisable matrix with no eigenvalues on the closed negative real line.

The most notable property which does not fully generalise from the scalar to the matrix exponential is that, in general, $\operatorname{expm}(A+B) \neq \operatorname{expm}(A)\operatorname{expm}(B)$. If A and B commute then the terms are equal, though their equality does not imply that the matrices must commute [7]:

$$AB = BA \Rightarrow \operatorname{expm}(A+B) = \operatorname{expm}(A)\operatorname{expm}(B) = \operatorname{expm}(B)\operatorname{expm}(A), \tag{B.12}$$

If A and B are diagonalisable and commute, then they are jointly diagonalisable [8], i.e. there exists some U which simultaneously gives $A = UD_AU^{-1}$ and $B = UD_BU^{-1}$. In this case, we have:

$$\begin{aligned} \operatorname{logm}(AB) &= \operatorname{logm}\left(U(D_AD_B)U^{-1}\right) = U \operatorname{logm}(D_AD_B) U^{-1} \\ &= \sum_{i=1}^{n} u_i \log(\lambda_i^A \lambda_i^B) v_i^T = \sum_{i=1}^{n} u_i \left(\log(\lambda_i^A) + \log(\lambda_i^B)\right) v_i^T \\ &= \sum_{i=1}^{n} u_i \log(\lambda_i^A) v_i^T + u_i \log(\lambda_i^B) v_i^T = \operatorname{logm}(A) + \operatorname{logm}(B). \end{aligned} \tag{B.13}$$

The special case of a matrix commuting with itself implies

$$\operatorname{expm}(kA) = (\operatorname{expm}(A))^k, \tag{B.14}$$

which can be used to show that even for non-diagonalisable A (B.11) holds (at least for integer powers):

$$\begin{aligned} \operatorname{expm}(k \operatorname{logm}(A)) = \operatorname{expm}\left(\sum_{i=1}^{k} \operatorname{logm}(A)\right) = \prod_{i=1}^{k} \operatorname{expm}(\operatorname{logm}(A)) = A^k \\ k \operatorname{logm}(A) = \operatorname{logm}\left(A^k\right). \end{aligned}$$

The special case of commutation between A and $-A$ gives

$$\operatorname{expm}(0_n) = \operatorname{expm}(A - A) = \operatorname{expm}(A)\operatorname{expm}(-A) = I_n. \tag{B.15}$$

A trivial case of (B.6) gives

$$\operatorname{logm}(kI) = I\log(k), \tag{B.16}$$

and then using (B.13) with the fact that kI commutes with any matrix gives

$$\operatorname{logm}(kA) = \operatorname{logm}(kI\,A) = \operatorname{logm}(kI) + \operatorname{logm}(A) = I\log(k) + \operatorname{logm}(A)\,. \tag{B.17}$$

Bibliography

[1] G. H. Golub and C. F. Van Loan, *Matrix computations*, 3rd ed. Johns Hopkins, 1996. ˆ369

[2] E. Denman and A. Beavers, "The matrix sign function and computations in systems," *Appl. Math. Comput*, vol. 2, no. 1, pp. 63–94, 1976. ˆ369

[3] S. M. Smith, N. D. Stefano, M. Jenkinson, and P. M. Matthews, "Normalized accurate measurement of longitudinal brain change." *J Comput Assist Tomogr*, vol. 25, no. 3, pp. 466–475, 2001. ˆ369

[4] C. Moler and C. Van Loan, "Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later," *SIAM Review*, vol. 45, no. 1, pp. 3–50, 2003. ˆ369

[5] N. Higham, *Functions of matrices: theory and computation*. Society for Industrial and Applied Mathematics, 2008. ˆ370

[6] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 29, no. 1, p. 328, 2008. ˆ370

[7] R. Horn and C. Johnson, *Topics in Matrix Analysis*. Cambridge University Press, 1994. ˆ371

[8] ——, *Matrix analysis*. Cambridge University Press, 1990. ˆ371

Appendix C

Procrustes Analysis

This short appendix presents results related to the problem of finding a matrix that best transforms one set of points into another, and the related problem of finding a constrained matrix that best approximates a more general one. The results have obvious application to point-based registration [1], but also relate to tensor reorientation [2] and to the removal of 'pose' in morphometric analysis [3].

C.1 Introduction

'Procrustes analysis', or the 'orthogonal Procrustes problem', refers to finding a matrix with orthonormal columns, (e.g. a rotation matrix, if square), which best relates two point-sets in terms of their summed squared error [4]. The 'extended orthogonal Procrustes' problem additionally allows (isotropic) scaling [5]. Generalised procrustes analysis can refer, for example, to finding a matrix with orthogonal, but not orthonormal columns [6], to the problem with more than two point-sets/matrices [7], or to weighting (also known as weighted Procrustes analysis) [8, 9].

The requisite mathematical tools, comprising some basic matrix calculus results and the method of Lagrange multipliers, are presented first, followed by the simple (unconstrained) case of finding the best (least-squares) affine transformation. Various cases of constrained transformations are then derived.

C.1.1 Selected results from matrix calculus

The only results needed for this appendix are two basic expressions for the derivatives of traces [10]. For ease of reference, they are presented with various forms arising from the invariance of the trace to transposition or cyclic permutation. In the following, A and B are matrices which do not depend on the matrix X.

$$\frac{\partial \mathrm{tr}\,(AX)}{\partial X} = \frac{\partial \mathrm{tr}\,(XA)}{\partial X} = \frac{\partial \mathrm{tr}\left(A^T X^T\right)}{\partial X} = \frac{\partial \mathrm{tr}\left(X^T A^T\right)}{\partial X} = A^T. \tag{C.1}$$

$$\begin{aligned}\frac{\partial \mathrm{tr}\left(AXBX^T\right)}{\partial X} &= \frac{\partial \mathrm{tr}\left(XBX^TA\right)}{\partial X} = \frac{\partial \mathrm{tr}\left(X^TAXB\right)}{\partial X} \\ &= AXB + A^TXB^T.\end{aligned} \tag{C.2}$$

Simplifications of (C.2) for the special cases of $A = I$ or $B = I$ are frequently of use:

$$\begin{aligned}\frac{\partial \mathrm{tr}\left(X^TAX\right)}{\partial X} &= (A + A^T)X, \\ \frac{\partial \mathrm{tr}\left(XBX^T\right)}{\partial X} &= X(B + B^T).\end{aligned}$$

Note that the first of these also gives the derivative of a quadratic form in the column vector x, since a scalar is equal to its own trace:

$$\frac{\partial x^TAx}{\partial x} = \frac{\partial \mathrm{tr}\left(x^TAx\right)}{\partial x} = (A + A^T)x.$$

C.1.2 Constrained optimisation

The method of Lagrange multipliers can be used to minimise a cost function while satisfying equality constraints [10].[1]

To minimise a scalar function of a matrix $\phi(X)$, subject to an $m \times p$ matrix function of equality constraints $G(X) = 0$, the Lagrangian is given by

$$L(X) = \phi(X) - \mathrm{tr}\left(\Lambda^TG(X)\right), \tag{C.3}$$

where Λ is an $m \times p$ matrix of unknown Lagrange multipliers.

Note that (C.1) gives

$$\frac{\partial L}{\partial \Lambda} = G(X),$$

i.e. zeroing the derivative with respect to the Lagrange multipliers recovers the constraint. Equating the derivative with respect to X to zero, gives a set of equations in X and Λ, which can be solved for X by employing the constraint to eliminate Λ.

C.2 Affine and linear transformations

Consider n pairs of approximately corresponding points $\{a_i, b_i\}_{i=1}^n$ each in $\mathbb{R}^m$. To find the affine transformation $(x \to Tx + t)$ that best relates the point-sets, it is necessary to minimise a suitable norm of the matrix E that collects together all of the column vectors

[1] Technically, the stationary points of the Lagrangian are not necessarily minima of the cost function. In this appendix, we assume that stationary points represent optimal solutions, though a more rigorous development would prove this in each case.

of correspondence error components:

$$\begin{aligned} Ta_i + t &= b_i + e_i \\ TA + t\underline{1}^T &= B + E, \end{aligned}$$

where $\underline{1}$ denotes the length n column vector of ones.[2]

It is common to minimise the root-mean-square of the correspondence error vector magnitudes, which is equivalent to the Frobenius norm of E. For simplicity, its square is minimised:

$$\|E\|_F^2 = \text{tr}\left(E^T E\right) = \text{tr}\left((TA + t\underline{1}^T - B)^T (TA + t\underline{1}^T - B)\right) \tag{C.4}$$

Considering first the translation:

$$\begin{aligned} \frac{\partial \|E\|_F^2}{t} &= 0 \\ &= \frac{\partial}{\partial t}\text{tr}\left(\underline{1}t^T t\underline{1}^T - 2\underline{1}t^T(TA - B)\right) \\ &= \frac{\partial}{\partial t}\text{tr}\left(t^T t\underline{1}^T\underline{1} - 2t^T(TA - B)\underline{1}\right) \\ &= 2nt - 2(TA - B)\underline{1} \\ t &= TA\underline{1}/n - B\underline{1}/n, \end{aligned}$$

which shows the intuitive result that the translation aligns the centroids of the point-sets TA and B, since e.g. $B\underline{1}/n$ simply averages B's columns.

For the linear part of the transformation

$$\begin{aligned} \frac{\partial \|E\|_F^2}{T} &= 0 \\ &= \frac{\partial}{\partial T}\text{tr}\left(A^T T^T TA - 2A^T T^T(t\underline{1}^T - B)\right) \\ &= \frac{\partial}{\partial T}\text{tr}\left(T^T TAA^T - 2T^T(t\underline{1}^T A^T - BA^T)\right) \\ &= 2TAA^T - 2t\underline{1}^T A^T + 2BA^T \\ &= 2TAA^T - \frac{2}{n}(TA - B)\underline{1}\underline{1}^T A^T + 2BA^T \\ TAA^T &= TA\underline{1}\,\underline{1}^T A^T/n - B\underline{1}\,\underline{1}^T A^T/n + BA^T \\ TA(I - \underline{1}\,\underline{1}^T/n)A^T &= B(I - \underline{1}\,\underline{1}^T/n)A^T \\ TA\bar{M}A^T &= B\bar{M}A^T, \end{aligned}$$

which has minimum-norm solution

$$T = B\bar{M}A^T(A\bar{M}A^T)^+,$$

where $\bar{M} = I - \underline{1}\,\underline{1}^T/n = I - \underline{1}\,\underline{1}^+$ is the projection matrix onto the null space of $\underline{1}$ or in other

[2] This is written as $1_{n\times 1}$ elsewhere in this thesis, but compactness is helpful in the derivations here.

words, the matrix which removes the mean column when used to post-multiply. Since this is a (symmetric and idempotent) projection matrix (see section A.1), one can split the products, e.g. $B\bar{M}A^T = B\bar{M}\bar{M}A^T = B\bar{M}(A\bar{M})^T$, meaning that the above expression is the usual least-squares solution after the centroids have been removed from the point-sets.

Considering $TA = B + E$ as a general matrix approximation problem (where A and B need not have the same number of rows, i.e. T may be rectangular) it is possible to swap the order of TA, i.e. $AT = B+E$. Since the Frobenius norm is unaffected by transposition, $\|TA - B\|_F$ can be written $\|A^TT^T - B^T\|_F = \|CF - D\|_F$, with solution

$$F = T^T = (A\bar{M}A^T)^+A\bar{M}B^T = (C^T\bar{M}C)^+C^T\bar{M}D,$$

which is easily recognised as the familiar least-squares solution, corresponding to the maximum-likelihood solution under a Gaussian assumption, as derived in section A.4.2.

C.3 Orthogonal and orthonormal transformations

If the matrix is restricted to have orthonormal columns, corresponding, in the square case, to a rotation (possibly with reflection, see below) then the method of Lagrange multipliers can be used to solve the constrained optimisation. Naturally, a translation would still align the centroids, and the matrix could be derived from the centred point-sets [3], so translation can be ignored henceforth.

Consider the case of a geometric similarity transformation; this can be written as fR where f is a scale-factor and R is a rotation, hence constrained to satisfy $R^TR = I$.[3] The more general case of orthogonal but not orthonormal columns (i.e. different scaling factors for each column) does not give rise to a closed-form solution, but numerical methods can be found in [6].

The Lagrangian is given by

$$L = \|fRA - B\|_F^2 + \operatorname{tr}\left(\Lambda^T(R^TR - I)\right), \tag{C.5}$$

where, as usual, zeroing of the partial derivative with respect to Λ recovers the constraint.

[3]The geometric similarity transformation is a scaled square rotation matrix fR, additionally satisfying $RR^T = I$, but the derivations here are valid for rectangular R with orthonormal columns.

Expanding the first term of the Lagrangian, and differentiating with respect to R:

$$\begin{aligned} L &= \text{tr}\left((fRA-B)^T(fRA-B)\right) + \text{tr}\left(\Lambda^T(R^TR-I)\right) \\ &= \text{tr}\left(B^TB\right) + f^2\text{tr}\left(A^TR^TRA\right) - 2f\text{tr}\left(A^TR^TB\right) + \text{tr}\left(\Lambda^T(R^TR-I)\right) \qquad \text{(C.6)} \\ \frac{\partial L}{\partial R} &= 0 \\ &= f^2\frac{\partial \text{tr}\left(RAA^TR^T\right)}{\partial R} - 2f\frac{\partial \text{tr}\left(R^TBA^T\right)}{\partial R} + \frac{\text{tr}\left(R\Lambda^TR^T\right)}{\partial R} \\ &= 2f^2R(AA^T) - 2fBA^T + R(\Lambda+\Lambda^T) \\ BA^T &= R\left(f^2AA^T + \frac{\Lambda+\Lambda^T}{2}\right). \end{aligned}$$

Now, using $R^TR = I$, and noting the symmetry of both AA^T and $\Lambda + \Lambda^T$

$$\begin{aligned} (BA^T)^T(BA^T) &= \left(f^2AA^T + \frac{\Lambda+\Lambda^T}{2}\right)^2 \\ &= (R^TBA^T)^2 \end{aligned}$$

and then using the compact singular value decomposition of $BA^T = USV^T$,

$$\begin{aligned} VS^2V^T &= (R^TUSV^T)^2 \\ VSV^T &= R^TUSV^T \\ V &= R^TU \\ UV^T &= UU^TR. \end{aligned}$$

It can now be observed that $R = UV^T$ is a sufficient solution. If BA^T is full rank, then the compact and full SVD coincide (see section A.2) implying $UU^T = I$ and hence that it is also a necessary solution.

At this point, note that the above result is independent of the value of f, including if it is set a priori to unity, and hence $R = UV^T$ is the optimal solution with orthonormal columns. Continuing with the orthogonal solution, differentiating the Lagrangian (C.6) with respect to f, before substituting in $R = UV^T$ and $BA^T = USV^T$

$$\begin{aligned} \frac{\partial L}{\partial f} &= 0 \\ &= 2f\text{tr}\left(A^TR^TRA\right) - 2\text{tr}\left(A^TR^TB\right) \\ f &= \frac{\text{tr}\left(A^TR^TB\right)}{\text{tr}\left(A^TR^TRA\right)} = \frac{\text{tr}\left(R^TBA^T\right)}{\text{tr}\left(AA^T\right)} \\ &= \frac{\text{tr}\left(VU^TUSV^T\right)}{\text{tr}\left(AA^T\right)} = \frac{\text{tr}\left(S\right)}{\text{tr}\left(AA^T\right)} \end{aligned}$$

As with the affine/linear case, the Procrustes problem is sometimes presented with the transformation on the right, i.e. $E = AT - B$, for which the relevant SVD becomes

$USV^T = A^TB$, after which $R = UV^T$ and $f = \mathrm{tr}\left(S\right)/\mathrm{tr}\left(AA^T\right)$ remain the same.

The above derivations find a matrix R satisfying $R^TR = I$, or $|R| = \pm 1$; it is often geometrically desirable to find a pure rotation without reflection, i.e. a special orthogonal matrix R with $|R| = 1$. From the matrix-approximation property of the singular value decomposition $BA^T = USV^T$, it is intuitively clear if the matrix $R = UV^T$ has negative determinant, it can be reflected with least additional squared error by replacing the positive unity element corresponding to the smallest singular value in an implicit identity matrix (made explicit in UIV^T) with -1. I.e. instead of $USV^T \rightarrow UV^T$, the approximation is $USV^T \rightarrow U\tilde{S}V^T$, where $\tilde{S}$ is a diagonal matrix with positve 1 everywhere except for the element corresponding to the final (smallest) element of S, which has -1. This result has been derived more rigorously by Umeyama [11].

C.3.1 Closest orthogonal matrix

An interesting special case of the objective function in equation (C.5) occurs if A is set to an identity matrix. Then, the problem is to approximate an arbitrary matrix B with a matrix fR satisfying $R^TR = I$. A special case of this problem is to find the closest rotation (or geometric similarity) to a general linear transformation. The solution is given by $R = UV^T$ and $f = \mathrm{tr}\left(S\right)/\mathrm{tr}\left(I\right)$, where $USV^T = B$ is the SVD of the linear transformation (potentially the linear part of a homogeneous affine transformation matrix) and f is the arithmetic mean of its singular values.

Bibliography

[1] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration." *Phys Med Biol*, vol. 46, no. 3, pp. R1–45, Mar. 2001. ˆ373

[2] D. Xu, X. Hao, R. Bansal, K. J. Plessen, and B. S. Peterson, "Seamless warping of diffusion tensor fields." *IEEE Trans. Med. Imag.*, vol. 27, no. 3, pp. 285–299, Mar. 2008. ˆ373

[3] F. L. Bookstein, "Landmark methods for forms without landmarks: morphometrics of group differences in outline shape." *Med Image Anal*, vol. 1, no. 3, pp. 225–243, Apr. 1997. ˆ373, 376

[4] P. Schönemann, "A generalized solution of the orthogonal Procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966. [Online]. Available: http://www.springerlink.com/content/l8023011278587w7/ ˆ373

[5] P. Schönemann and R. Carroll, "Fitting one matrix to another under choice of a central dilation and a rigid motion," *Psychometrika*, vol. 35, no. 2, pp. 245–255, 1970. [Online]. Available: http://www.springerlink.com/content/gq1933g700170w46/ ˆ373

[6] R. Everson, "Orthogonal, but not orthonormal, Procrustes problems," 1998, submitted to Advances in Computational Mathematics. ˆ373, 376

[7] J. Gower, "Generalized Procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975. ˆ373

[8] R. Lissitz, P. Schönemann, and J. Lingoes, "A solution to the weighted Procrustes problem in which the transformation is in agreement with the loss function," *Psychometrika*, vol. 41, no. 4, pp. 547–550, 1976. [Online]. Available: http://www.springerlink.com/content/x03267656316n0l5/ ˆ373

[9] P. Batchelor and J. M. Fitzpatrick, "A study of the anisotropically weighted Procrustes problem," in *Proc. IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, 2000, pp. 212–218. ˆ373

[10] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 2nd ed. John Wiley, 1999. [Online]. Available: http://cdata4.uvt.nl/websitefiles/magnus/mdc.pdf ˆ373, 374

[11] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 4, pp. 376–380, 1991. ˆ378

Appendix D

Permutation test implementation

Chapter 2 contains theoretical derivations of various permutation-testing strategies for general linear models, and chapter 4 proposes two more specialised test statistics for application to multivariate two-sample problems on general data and on directional data. In this appendix, we discuss practical computational issues in the implementation of these methods. All of this work is novel in the sense of being done from scratch, but most of the software remains relatively simple, so no great novelty in implementation can be claimed. Nevertheless, this appendix contains details that are often omitted from published papers, so should be of use to others attempting to implement permutation tests for imaging or other computationally-demanding applications.

D.1 Introduction

For large imaging data-sets it is crucial that the implementation is efficient in terms of both memory usage and computational speed. In some circumstances, the complete data matrix (with dimension $n \times m \times v$ for n subjects, m multivariate measurements per voxel, and v voxels) could be too large to fit into the available memory. A basic implementation would result in the operating system swapping data to and from disk, which is unlikely to occur in an optimal way, hence degrading computational speed. In particular, if we wish to loop over voxels within an outer loop over permutations (see below) swapping would be catastrophic for performance. A solution to this will be presented below.

High-performance computing clusters are becoming commonplace in academia and industry, allowing considerable reductions in overall computation time if code can be distributed to run in parallel across multiple processing cores. The large numbers of both voxels and desired permutations immediately suggests such parallelisation should be possible, but care must be taken to distribute data and processing tasks in the correct way, and this will be addressed in D.5.

We first comment very briefly on the relative efficiency of the different general linear model permutation methods presented in chapter 2, and outline a general way to make the computation of the test statistic less costly.

D.2 Efficiency for general linear model permutation

For voxel-wise nuisance covariates, ter Braak's method is more computationally efficient than Freedman-Lane, since different contrasts (which alter the reduced- but not the full-model residuals) can be computed simultaneously. The Shuffle-Z method is perhaps better still, since it doesn't require that the data be orthogonalised with respect to the (varying) nuisance. Importantly, note that while the Shuffle-Z like reformulation of Huh-Jhun may appear almost as efficient as Shuffle-Z, it suffers slightly with voxel-wise nuisance, since a new U_0 transformation matrix must be computed for each voxel. In practice, this probably necessitates the inner-permutation implementation (see section D.4 below) for Huh-Jhun with voxel-wise covariates, while Shuffle-Z could use either implementation.

D.2.1 Effect of permutation on projection matrices

The Wilks' Λ statistic (and also the less general F- or t-statistics) require the computation of perpendicular projection matrices. We presented computationally efficient formulations in terms of the singular value decomposition at the end of appendix A.3, of the form $P = UU^T$ and and $R = U_n U_n^T$, where U and U_n are respectively orthonormal bases for the column-space and left-null-space. The t-statistic additionally requires the estimated parameters that are obtained via the pseudo-inverse of the design matrix $X^+ = V\Sigma^{-1}U^T$.[1]

With large imaging data-sets it is undesirable to actually permute the observed data, so one would naturally permute the rows of the design matrix, and then recompute the singular value decomposition and pseudo-inverse. In fact, one can do this much more efficiently. The key is to begin by considering permutation of the data, so as to see the effect not on the design itself, but on the derived projection matrices or pseudo-inverse. For example, under permutation of the data,

$$\hat{B}_S = X^+ SY = V\Sigma^{-1}U^T SY,$$

and it is immediately obvious that we can simply permute the columns of U^T and avoid re-computing the singular value decomposition or pseudo-inverse. Similarly for the quadratic forms that occur in the conventional statistics, we have, for example,

$$(SY)^T R(SY) = Y^T S^T U_n U_n^T SY,$$

meaning that we can permute just the precomputed $n \times (n-r)$ matrix U_n compute the $(n-r) \times m$ term $U_n^T SY$, and then form the $m \times m$ matrix of sums of squares and cross-products very efficiently.

D.3 Blocking

To illustrate the need for blocking, consider the following large but realistic data-set. Whole-brain images resampled in the space of the 2 mm isotropic MNI/ICBM 152 template

[1] Here, we use Σ in the SVD, and S for a 'shuffling' or permutation matrix.

have $91 \times 109 \times 91 = 902629$ voxels; if each voxel contains a complete 3×3 Jacobian matrix, and calculations are performed in double-precision, then 66 scans exceed the 4GB limit addressable with 32-bit architectures. Higher resolution may be desirable, particularly with searchlight-mapping of unmoothed data (see Chapter 4); with 1 mm isotropic data, the above example would require 32 GB, which is far from universally available even on modern high-performance cluster nodes. Sixty subjects may seem large in comparison to the average published neuroimaging experiment, but there is in fact a trend towards even larger structural MRI projects, for example over 400 subjects are available from the OASIS project [1], while the Alzheimer's Disease Neuroimaging Initiative [2] has recruited 800.

There are virtually limitless ways in which large data-sets could be split into blocks of a certain size, but the layout of voxels within memory means that reading one or more planes together will result in faster contiguous memory access. The current version of SPM (`http://www.fil.ion.ucl.ac.uk/spm/software/spm5/`) performs its statistics with a single plane (or potentially even smaller blocks per plane) in memory at a time, but this is very conservative.

While most voxel-wise statistical analyses could be performed with a simple strategy of loading neighbouring collections of planes, the searchlight (4) requires access to all voxels within a certain spherical neighborhood around each voxel. This means firstly that including data from multiple planes is essential, even if it means that entire planes cannot be read at once for very large data-sets, and secondly that each block of data must overlap the adjacent block(s) by the radius of the neighborhood in voxels. We have implemented a general blocking strategy, favouring planar blocks but accounting for overlap if necessary, which can access multiple three, four or five dimensional NIfTI volumes, using the SPM5 `@nifti` class. Blocks are permuted so that scan and variable indices are faster changing than voxel-indices, as this should improve performance of multiple voxel-wise statistical tests. Voxel data (possibly multivariate) within a specified spherical neighborhood are automatically combined into (larger) multivariate observations, with standard voxel-wise analysis then simply becoming the special case of a zero-radius neighborhood.

D.3.1 Relation to FWE and step-down procedures

Use of the maximum-distribution for family-wise error control [3] clearly interacts with the need for blocking; the distribution of interest is the maximum from anywhere in the image, not simply the currently considered block. This is trivial to deal with though, since the maximum over the maxima of the separate blocks is clearly the overall maximum, so it is only necessary to record separate permutation distributions of the maximum-statistic in each block, and take the maximum over this when all permutations and blocks have been completed.

The step-down FWE procedure of Belmonte and Yurgelun-Todd [4] discussed in section 2.3.1 complicates this slightly, since it requires a number of 'secondary maxima' to be recorded. It would not be correct to extend the above approach to first compute the overall maximum by merging the maxima from the blocks and then the overall secondary

maximum by merging the secondary maxima, since one of the secondary maxima from one block could be larger than the primary maximum from another. However, a little thought reveals that if we record the top N_r statistics for every block, these must jointly contain the top N_r statistics for the whole image. So an efficient step-down procedure is straightforward even with very large data-sets.

Sorted lists of secondary maxima should ideally be maintained in sophisticated data-structures, like the double-keyed binary tree structure used by Belmonte and Yurgelun-Todd [4], which allows both insertion of statistic values and later removal of voxel locations to be performed in logarithmic time. However, such pointer-based structures are not easily available within MATLAB. We observe though, that if each voxel's statistic is simply swapped with the current minimum value in a simple vector data-structure when it exceeds this minimum, then on completion, the vector will contain the correct set of values for each permutation, albeit not in the correct order. However, outside of the parallelisation and other loops, these vectors can simply be sorted as a post-processing step, independent of the number of blocks or nodes, which provides a reasonably efficient and easily implemented compromise.

D.4 Looping over permutations and voxels

It may initially appear arbitrary whether one loops over all of the voxels in the mask, considering for each one a loop over the permutations, or vice versa. However, the order of the loops can have a significant effect on the the computational efficiency. Interestingly, the generic Wilks' Λ statistic and the two special-purposes statistics employed in chater 4 are suited to different choices here.

For Wilks' Λ, even with the computational simplifications presented in section D.2.1 it is still more efficient to compute the statistics on all the voxels (or at least, all within the current block) with the same permuted design, since the permuted basis matrices are then only computed N_p times instead of $N_p \times N_v$, and the multiplications of these matrices with the data are performed the same number of times.

However, regarding the Cramér statistic in the permutation framework, we note that relabelling of the group 1 and group 2 observations doesn't change the actual set of inter-point distances computed, it only selects in which summations they appear. This means that for a particular voxel, with the multivariate observations collected into the n-by-m matrix Y, we can precompute a matrix of all the kernelised distance values in ϕ. In MATLAB code:

```
K = Y*Y'; % inner-products
S = diag(K)*ones(1, n1+n2); % squared norm of each vector, replicated
D = S + S' - 2 * K; % squared distances between all observations
Phi = phi(D); % matrix Phi from (element-wise) function phi(z)
```

Following that, for any permutation of the vector `g` of boolean indicator variables denoting memebership of group $\mathcal{G}_1$ (with the rest of the observations in $\mathcal{G}_2$ being indicated by the Boolean complement `~g`, we can evaluate the Cramér statistic by:

```
GX = sum(sum(Phi(g,  ~g))) / (n1*n2);
G1 = sum(sum(Phi(g,   g))) / (n1^2);
G2 = sum(sum(Phi(~g, ~g))) / (n2^2);
T = (n1*n2 / (n1+n2)) * (2 * GX - G1 - G2);
```

Note that the first term of `T` can be dropped for permutation-testing of the Cramér statistic, as the group numbers are constant over both permutations and voxels. It is clear from this that, in contrast to Wilks' Λ above, the Cramér test is more efficient if the permutation loop occurs within the data loop, since then the matrix `Phi` need only be computed once per voxel.

Finally, the Watson statistic, presented in section 4.3.3 for comparing the principal axes of strain, requires computation of the eigenvalues of Y^TY, $Y_1^TY_1$ and $Y_2^TY_2$. For the larger of the two groups (assumed to be group 1, without loss of generality), we may compute

$$S_1 = Y_1^TY_1 = Y^TY - Y_2^TY_2,$$

i.e. a subtraction of two 3-by-3 matrices: the permutationally invariant Y^TY term which can be precomputed, and the $Y_2^TY_2$ term for the smaller group, which is needed anyway. This suggests that permuting within the data loop, as for the Cramér statistic, will be more efficient, since Y^TY is constant over permutations for each voxel, while both Y^TY and $Y_1^TY_1$ change over voxels. Note that the Cramér and Watson tests form respectively the $n \times n$ matrix YY^T and the $m \times m$ (for $m = 3$ typically) matrix Y^TY. It is of academic interest to note that the non-zero (and hence maximal) eigenvalues of these two matrices are in fact the same — they are the singular values of Y (see section A.2.2). So, instead of repeatedly computing $Y_1^TY_1$ for every permutation, using it to get $Y_2^TY_2$, and then computing the eigenvalues of both of these matrices, one could compute the largest singular value for each of Y_1 and Y_2. Alternatively, one could compute YY^T as for the Cramér statistic, and then compute the maximal eigenvalues of its groups' partitions. However, in practice, these approaches would probably not improve upon the simpler implementation, since they require eigen- or singular-value decompositions of larger matrices, which are typically more expensive than the matrix multiplications that they save.

D.5 Parallelisation

Given the large numbers of voxels and of permutations, for which essentially the same statistical test must be performed, there is obvious potential to exploit parallel computing architectures. There is also clearly a choice as to whether to allocate parallel tasks over voxel-space or design-space. With a simple permutation-test from which only uncorrected inference is required, it would be trivial to do either or even both. However, with a maximum-statistic based FWE procedure, it becomes much less efficient to distribute voxels (or planes or other sub-blocks) since each parallel task would need to contribute to the overall record of each permutation's image-wise maximum. With the step-down procedure employed here, this would essentially mean allowing all the parallel units to write anywhere within the same shared-memory data-structure for the maxima and the

reserves. Some form of locking would be required to prevent race conditions: consider one task swapping an existing maximum (or secondary reserve maximum) in the array for one that it has just found, while another task simultaneously wishes to do likewise; it would be possible for the second task to perform its swap in the time between the first task checking that its maximum value is larger than the one originally in the list and actually performing its swap, hence the first task could incorrectly move the potentially larger maximum from the second. This complication, and the likely reduction in speed caused by a suitable locking mechanism, essentially precludes the use of this form of parallelisation. Note though, that for efficient computation of uncorrected p-values, the complete statistic image for the original (identity permutation) labelling must be computed before any of the permutations are performed, so that for the permutations one need only count the numbers of times the original statistic is exceeded for each voxel (and keep track of maxima) instead of having to record the entire permutation distribution. For this identity-permutation statistic image, all voxel values are kept, hence there are no complications related to tracking the maximum statistic and its reserves, so it is trivial to distribute voxels or blocks of data here. With small data-sets the overhead in distributing parallel tasks might outweigh the benefits from parallelisation, but with large numbers of high-resolution images, there can be a noticeable speed-up.

Distributing permutations in parallel is relatively straightforward. Each parallel node keeps a record of the maxima over its set of permutations, as well as a count of the times the original statistics are exceeded. When all nodes have finished, their arrays of maximum-statistic information are simply concatenated in the permutations dimension, and their exceeded-counts are summed together. There is one slight complication, which is the interaction between the blocking strategy (required if each individual node has insufficient memory to hold the complete data, regardless of the amount of memory available in total in the computing cluster) and the distribution of permutations. Multiple blocks cannot run in parallel with the same set of permutations, since they would require simulatenous access to the same part of the recorded maximum-distribution. To save each node reading the data from disk, and performing the blocking themselves, we use one node to control the blocking in an outer loop, broadcasting (with MATLAB's `labBroadcast`) each block to all nodes, which then complete their sets of permutations. All nodes are then synchronised (using `labBarrier`) before the next block is broadcast.

D.6 Validation

It is clearly important that newly developed software is validated carefully, and this is particularly true in situations such as statistical testing, where small errors could produce results which are superficially apparently reasonable, and yet statistically invalid and potentially misleading. The basic Wilks' Λ and Cramér statistics have been checked using other independent software implementations. To be precise, only various special cases of our general Wilks' Λ code are easily available as MANOVA tests in MATLAB's statistical toolbox, but in addition, more general designs and contrasts have been evaluated in the

special case of univariate F-statistics with SPM. Between these two verifications it seems reasonable to conclude that the basic statistics are correct.

Validation of the multivariate permutation test is more difficult, due mainly to its greater novelty. For the univariate case, we tested that it replicates the results of FSL's randomise (`http://www.fmrib.ox.ac.uk/fsl/randomise/index.html`) with regard to the voxel-level uncorrected and FWE corrected p-values, on data-sets that are small enough to permit exhaustive permutation. At present, neither non-exhaustive permutation, nor the step-down FWE correction procedure have been validated. The former is difficult to test experimentally, so might best be verified from careful inspection of the code; the requirement is that there are no mistakes or biases in the generation of random design matrices.

One of the simplest but perhaps most important means for validating the software, is to compare the results of the parallel code, with blocking, searchlight, etc. to simpler implementations on data where the aforementioned features are unnecessary. This has been done as the software progressed, for example checking that the parallel implementation reproduced the results of an earlier non-parallel version.

D.7 Further work

The most obvious potential improvement of the algorithms would be reimplementation in a programming language better suited than MATLAB to the task of iterating over a large number of permutations and voxels. A language with efficient pointer-based binary-tree structures such as C would be ideal. Our current software runs in parallel on a dedicated computing cluster; an appealing alternative would be a general purpose graphics processing unit implementation, such as that developed in CUDA[2] for non-rigid registration by Modat et al. [5].

Bibliography

[1] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults." *J Cogn Neurosci*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007. ^382

[2] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "The Alzheimer's disease neuroimaging initiative." *Neuroimaging Clin N Am*, vol. 15, no. 4, pp. 869–77, xi–xii, Nov. 2005. ^382

[3] T. Nichols and S. Hayasaka, "Controlling the familywise error rate in functional neuroimaging: a comparative review." *Stat Methods Med Res*, vol. 12, no. 5, pp. 419–446, Oct. 2003. ^382

[2] `http://www.nvidia.com/object/cuda_home.html`

[4] M. Belmonte and D. Yurgelun-Todd, "Permutation testing made practical for functional magnetic resonance image analysis," *IEEE Trans. Med. Imag.*, vol. 20, no. 3, pp. 243–248, Mar. 2001. ˆ382, 383

[5] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin, "Fast free-form deformation using graphics processing units." *Comput Methods Programs Biomed*, 2009, in press. [Online]. Available: http://eprints.ucl.ac.uk/17431 ˆ386